UNIVERSITÉ DE SHERBROOKE
Faculté de génie
Département de génie civil

MODÈLE STATISTIQUE DE PRÉVISION LONG TERME DE PRODUCTION HYDRO-ÉLECTRIQUE

Mémoire de maîtrise
Spécialité : génie civil

Pierre-Olivier CARON-PÉRIGNY

Sherbrooke (Québec) Canada

Avril 2020

MEMBRES DU JURY

Robert LECONTE
(Directeur)

Mélanie TRUDEL
(Évaluatrice)

Guillaume TAREL
(Évaluateur externe)

RÉSUMÉ

Le manque de données est un des facteurs limitatifs à l'implantation de modèles fiables permettant la prévision de la production d'hydro-électricité chez certains producteurs. En effet, la capacité de prévision des modèles conventionnels conceptuels ou à base physique dépend entièrement de la qualité et de la quantité des données physiques disponibles pour analyses. Souvent, à qualité de données égale, le gain en précision d'un modèle est directement relié à sa complexité. Passé un certain stade, les bénéfices associés par ce gain ne justifient plus les investissements nécessaires à l'instrumentation supplémentaire d'un bassin versant. De plus, la qualité des données utilisées pour la modélisation peut être remise en cause par des erreurs de mesure ou d'échantillonnage. Souvent, les appareils de mesures sont imprécis ou se situent dans des endroits éloignés occasionnant des coûts importants pour leur entretien et la collecte de données. Afin de limiter les coûts d'instrumentation supplémentaires et de réduire les risques liés à l'utilisation de données erronées, les données disponibles et fiables doivent être exploitées à leur juste valeur. Ce projet de recherche propose d'utiliser un minimum de données fiables, soit la production historique journalière de plus d'une centaine de centrales hydro-électriques sur près de 40 ans, pour développer un modèle statistique, non à base physique, de prévision de production hydro-électrique. Ce modèle propose l'utilisation de plusieurs techniques statistiques avancées ainsi que des algorithmes de forage de données telle que la déformation temporelle dynamique. Ce modèle de prévision a permis d'améliorer la précision sur tous les horizons testés en comparaison de la moyenne à long terme (LTA). Bien que l'objectif principal du projet soit d'améliorer la précision des prévisions de production d'hydro-électricité, son utilisation permet aussi de générer des résultats de manière probabiliste, permettant ainsi de communiquer l'incertitude de la prévision et d'aider à la prise de décisions basée sur ces résultats.

Mots-clés : Prévision de production hydro-électrique; Modèle statistique; Déformation temporelle dynamique

REMERCIEMENTS

Je tiens à remercier toutes les personnes qui m'ont aidé lors de la rédaction de ce mémoire. Plus particulièrement, je voudrais remercier Énergie Brookfield d'avoir cru en ce projet et Robert Leconte de m'avoir épaulé tout au long de ce parcours même après le sempiternel *Oui, oui ça avance!*

Merci également à mes amis et collègues Richard, Jean-Guillaume, Mathieu, Bruno, Colin et Jasmin. Sans leur support, ce travail serait rapidement devenu douloureusement pénible. Sans oublier mes parents Claire et André et surtout ma copine Joëlle qui a subi plus que quiconque les soirées/nuits de travail et le stress encouru.

Finalement, sans la remercier, une certaine tumeur m'a bien fait comprendre que la vie ce n'est pas fait pour abandonner.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

1 INTRODUCTION

1.1 Mise en contexte et problématique

Les problèmes environnementaux causés par les changements climatiques et, conséquemment, les nouveaux protocoles internationaux destinés à réduire les émissions de gaz à effet de serre ont conduit à une attention accrue aux sources de production durable d'électricité. Le secteur de l'hydro-électricité, déjà bien ancré au Québec est un acteur important du mix énergétique nord-américain. Les anciennes et les nouvelles centrales hydro-électriques font l'objet d'études de faisabilité et d'évaluations économiques de plus en plus approfondies. Contrairement aux centrales thermiques qui tirent leur énergie de carburants comme le gaz, le charbon ou le pétrole, dont l'approvisionnement dépend principalement d'activités humaines, l'hydro-électricité est presque entièrement soumise aux cycles hydrologiques naturels.

Le manque de données de qualité est l'un des facteurs les plus limitatifs à la mise en place de modèles fiables de prédiction de la production hydro-électrique. En effet, la capacité de prévision des modèles classiques basés sur la simulation du cycle hydrologique dépend grandement de la qualité et la quantité des données physiques amassées et disponibles pour l'analyse. L'approche habituelle d'un modèle de prévision de la production d'hydro-électricité est de décortiquer l'ensemble du système hydrologique en une suite de plusieurs sous-processus. Chacun d'eux est estimé par une expression mathématique. Le résultat de chaque estimation devient une

des données d'entrée pour le maillon suivant de la chaîne de modélisation, jusqu'à ce que le problème soit résolu. Tel que montré à la figure 1.1, pour un la portion hydrologique seulement d'un bassin versant, les sous-processus peuvent inclure par exemple: des modèles représentant la chute de neige, les changements de température, l'évolution de l'albédo, de l'évaporation ou de la vitesse d'infiltration et de ruissellement.

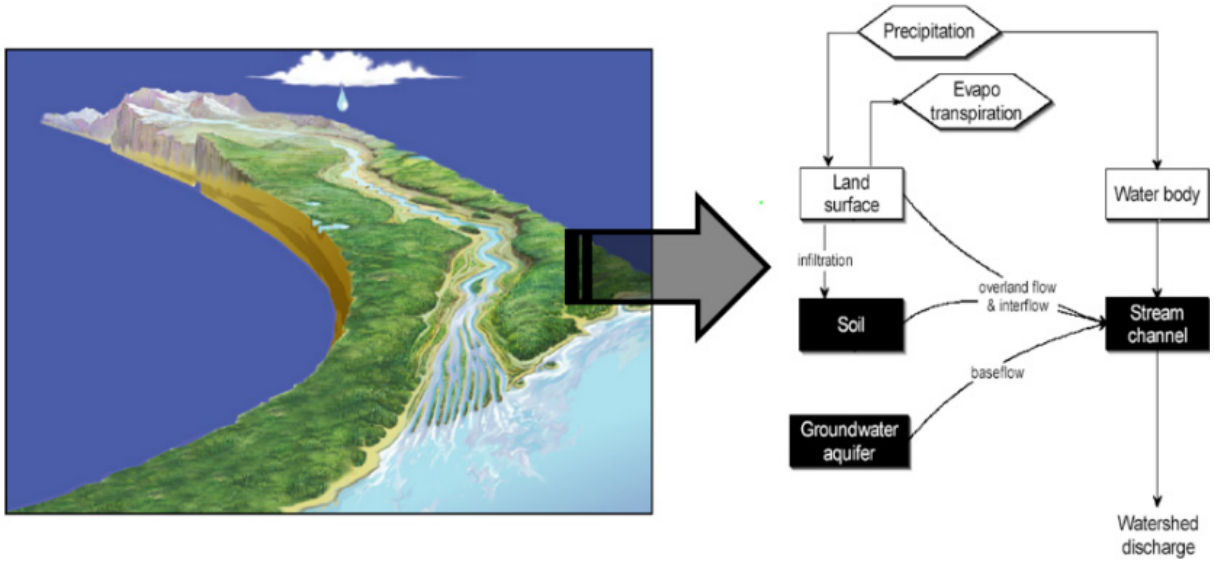


Figure 1.1 Exemple de modèle hydrologique à base physique [33]

Souvent, le gain en précision d'un modèle est généralement lié à sa complexité [7, 23]. Passé un certain stade, les avantages associés à la complexification du modèle ne justifient plus l'investissement requis pour l'instrumentation supplémentaire d'un bassin versant. En raison de leur nature même, les dispositifs de mesure, pour offrir une bonne répartition sur le bassin versant, sont de façon générale, placés dans des endroits éloignés et difficiles d'accès. Ces points de mesure ont pour effet de rendre les campagnes d'échantillonnage et de calibration des exercices rares, complexes et coûteux. Par conséquent, si pour réduire les coûts d'instrumentation, les points de mesure sont positionnés en fonction de la

facilité d'accès aux sites, la qualité et surtout la représentativité des données utilisées pour la modélisation peuvent être touchées parce que les mesures prises sur le terrain ne représentent pas conformément la réalité. Cette relation conceptuelle entre la complexité d'un modèle nécessitant beaucoup de données et les conséquences découlant de la qualité de ses prévisions est montrée à la figure 1.2.

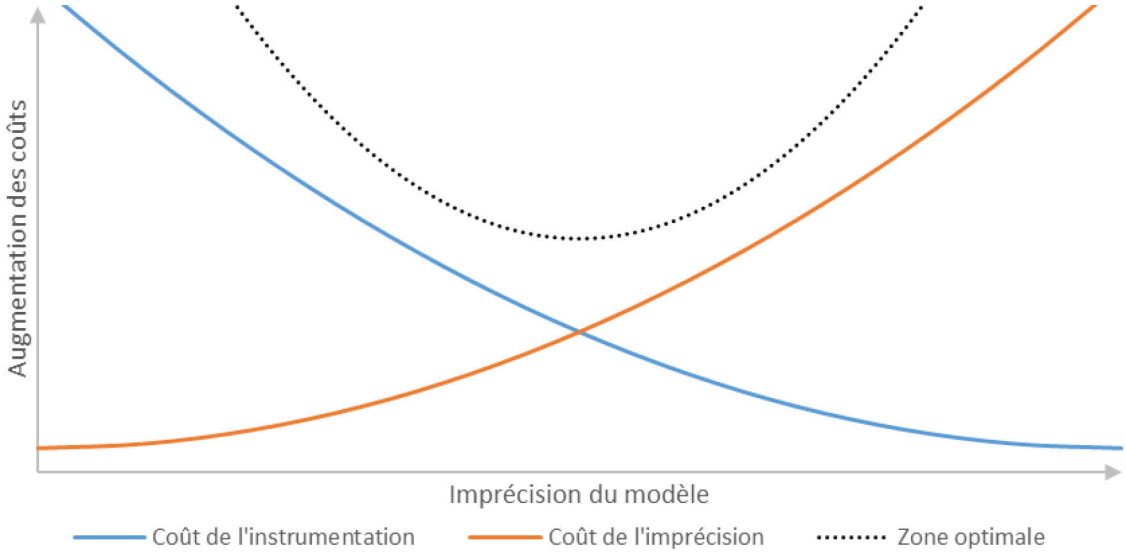


Figure 1.2 Coût d'utilisation d'un modèle prédictif [7]

Les coûts d'instrumentation supplémentaires et les risques associés à l'utilisation des données incorrectes engendrent de grands défis pour plusieurs producteurs canadiens. Ainsi, les prévisions de production hydro-électrique d'un horizon de plus de quelques semaines sont souvent limitées à la moyenne historique de la production d'énergie, *long term average* (LTA), à cet endroit. En dépit d'être simple, lorsqu'utilisé pour des prévisions à long terme, le LTA ne tient pas compte de l'état actuel de la génération, ce qui peut causer des attentes de production d'hydro-électricité souvent

irréalistes. Utilisé comme tel, pour un domaine aussi complexe que l'hydrologie, le LTA de la production historique est rarement un bon indicateur de la génération d'énergie future. La problématique du LTA est que cette méthode ne tire pas avantage de la plupart des informations contenues dans les données brutes.

Pour pallier à ces problèmes, grâce au développement de la puissance de calcul des ordinateurs et des avancées technologiques en matière de gestion de base de données, diverses techniques d'analyse et de forage de données (*Data mining*) sont utilisées pour extraire un maximum d'information des séries temporelles de données [9, 42]. Le forage de données est défini comme étant l'extraction d'informations implicites et non triviales, potentiellement utiles et précédemment inconnues à partir d'une base de données [19]. Le forage de données est utilisé pour découvrir des modèles et des patrons auparavant inconnus à partir de grands ensembles de données. Ainsi, le forage de données ne se limite pas à la collecte et à la gestion des données, mais comprend également l'analyse des structures de données et leur prédiction. Les outils développés pour ces fonctions proviennent de modèles statistiques, d'algorithmes mathématiques et de l'apprentissage automatique, tels que les réseaux de neurones et les arbres de décision. Plus particulièrement, en hydrologie et, implicitement, dans le domaine de la génération hydro-électrique, le forage de données est appliqué à la recherche de signes indicateurs d'inondations [31], à l'élaboration de règles de gestion de réservoir [47], à la prévision de ruissellement de surface [44].

1.2 Question de recherche

Tel que mentionné à la section précédente, les approches de forage de données ont été appliquées pour solutionner divers problèmes reliés à l'hydrologie. Toutefois, leur application/adaptation pour la prévision de la production hydro-électrique par l'utilisation de séries historiques de production reste à être démontrée. Ainsi, la question de recherche qui se pose est la suivante :

Dans quelle mesure l'utilisation de techniques de forage de données permet de faire la prévision de la production hydro-électrique de systèmes hydriques possédant des caractéristiques techniques, physiques, géographiques et climatiques variées?

Ainsi, cette recherche propose d'adapter des techniques statistiques avancées ainsi que des algorithmes de forage de données pour créer un outil de prévision de production hydro-électrique journalière d'un horizon allant de 10 à 150 jours. Cet outil n'utilise que les données de production hydro-électrique journalières historiques comme intrant et ne nécessite donc aucune instrumentation supplémentaire sur le terrain, réduisant ainsi l'incertitude liée à la qualité et la quantité de données hydrologiques disponibles. Ce dernier point est particulièrement pertinent pour les petits producteurs ou ceux exploitant plusieurs sites éloignés géographiquement qui ne disposent pas nécessairement des ressources nécessaires pour établir un modèle de prévision sophistiqué.

De plus, l'utilisation d'un outil statistique de prévision, d'une complexité supérieure au LTA, permettrait de suivre une tendance nouvelle dans la gestion de la production hydro-électrique et de passer de prévisions déterministes à probabilistes. Cette approche probabiliste, longtemps étudiée par le domaine universitaire, gagne du terrain dans les secteurs opérationnels publics et privés de l'industrie hydro-électrique [13]. Par définition, l'approche déterministe fournit un scénario unique qui nie le fait que plusieurs sources d'incertitudes influencent la prévision. Par conséquent, les décisions concernant la gestion de réservoirs, la commercialisation de l'énergie et la planification financière ne prennent en considération que ce seul scénario. L'incertitude liée à cette prévision n'est pas considérée, ce qui peut potentiellement entraîner une mauvaise décision. Aussi, les scénarios hors normes, improbables, mais possibles, ne sont généralement pas communiqués, malgré le risque qu'ils peuvent poser à la sécurité humaine, matérielle et financière. Sans être utilisés pour la prévision de production directement, ils peuvent se montrer fort utiles pour la gestion de risque.

L'approche statistique mise de l'avant dans ce projet a donc aussi pour but de communiquer l'incertitude sur les prévisions de production hydro-électrique. Au lieu de produire une seule valeur comme prévision, un poids statistique est attribué aux scénarios produits pour en déduire les probabilités de dépassement. Sans réduire l'incertitude, cette méthode permet à tout le moins de la quantifier, donc de la communiquer et d'aider à la prise de décisions.

1.3 Objectifs du projet de recherche

L'objectif principal de cette recherche est de développer un outil de prévision de production hydro-électrique journalière basé sur des algorithmes de forage de données et techniques statistiques avancées et n'employant que des données de production hydro-électrique journalière comme intrant. L'horizon de prévision varie de 10 à 150 jours.

Afin de limiter les coûts d'instrumentations supplémentaires et de réduire les risques liés à l'utilisation de données erronées, les données disponibles et fiables doivent être exploitées au maximum. Ce projet de recherche propose d'utiliser un minimum de sources de données, les plus fiables possibles, soit la production historique journalière de plus d'une centaine de centrales hydro-électriques sur près de 40 ans. L'utilisation des données de production hydro-électrique directement, plutôt que les apports en eau à la centrale, permet de produire une prévision avec exactement l'unité de mesure finale. De plus, l'utilisation de ces données permet de borner les statistiques de distribution entre une production nulle et une production maximale théorique. Enfin, l'étude a pour objectif secondaire de transmettre les incertitudes reliées à la prévision de production aux preneurs de décisions par une fonction de distribution.

1.4 Plan du mémoire

Le chapitre 2 présente une revue de littérature, assurant la pertinence et le positionnement du projet en cours par rapport à l'état actuel de la recherche dans le domaine. L'état de l'art porte d'abord sur l'incertitude intrinsèque à

la modélisation. Avant de montrer les forces et les faiblesses de différentes familles de modèles et l'étendue de l'utilisation des statistiques en prévision hydrologique et hydro-électrique. Le chapitre 3 définit la méthodologie utilisée au sein du modèle statistique et la procédure afin d'évaluer la performance du modèle et la sensibilité des différents paramètres choisis. Les résultats sont montrés au chapitre 4 et sont accompagnés d'une discussion sur les étapes de calibration. Le document se termine au chapitre 5 avec une conclusion et une ouverture sur les potentiels futurs travaux utiles en lien avec ce projet de recherche.

2 ÉTAT DE L'ART

Ce chapitre différencie le risque de l'incertitude et commente certaines méthodes reliées au traitement et à la prévision de séries temporelles pour permettre de mieux comprendre le choix de la méthodologie proposée.

2.1 L'incertitude et le risque associés aux modèles de prévision

Le concept d'incertitude associée aux prévisions de production hydro-électrique peut être directement lié à deux types de composantes du modèle, soit à sa structure même et aux données qui y sont utilisées [69]. Tel que mentionné en introduction, la représentation conceptuelle d'un bassin versant traduite sous forme d'équations mathématiques dans un modèle est basée sur une compréhension limitée du système hydrologique réel [52]. Pour les modèles basés sur les données, ne cherchant pas à reproduire des phénomènes physiques, l'incertitude est principalement relative à l'estimation des paramètres [49]. Les erreurs d'évaluations de paramètres augmentent de façon inversement proportionnelle à la taille des jeux de données utilisés pour l'analyse et diminuent avec le nombre de paramètres du modèle. L'enjeu est de définir un jeu optimal de paramètres d'un modèle sur la base des informations disponibles. Il est bien connu que les modèles hydrologiques souffrent du problème d'équifinalité, suivant lequel plusieurs jeux de paramètres peuvent donner des simulations jugées satisfaisantes [4]. Plusieurs méthodes existent pour quantifier cette incertitude, telles que la *Generalised Likelihood Uncertainty Estimation*

(GLUE) [5] ou le calcul bayésien approximatif (ABC) [1]. Un autre élément d'incertitude dans le traitement des données est introduit lorsqu'un modèle est requis pour interpréter la mesure réelle. Un exemple typique est l'utilisation des mesures radar des précipitations. Ce sont des mesures de réflectivité qui doivent être transformées en estimations de précipitations utilisant un modèle (empirique) avec une relation fonctionnelle choisie et des paramètres calibrés, les deux pouvant être incertains.

Par définition, le risque est l'exposition à un événement indésirable dans un environnement incertain, comme le risque qu'un barrage ne cède. La probabilité est une mesure de la possibilité que l'événement indésirable n'arrive.

Les compagnies de production d'électricité tentent, entre autres, de gérer certains risques, tels que le risque lié aux fluctuations de prix et le risque lié aux difficultés d'approvisionnement. Ce dernier, pour les générateurs thermiques au charbon ou au gaz, par exemple, est principalement lié au coût et à la disponibilité du combustible. Les producteurs hydro-électriques quant à eux, doivent faire face aux aléas de la nature et ce risque est plutôt associé aux apports en eau aux différentes centrales. Ce risque, tout comme le risque lié à l'incertitude sur les prix futurs, peut être mitigé par l'utilisation de certains outils de couverture tels que les contrats à terme sur l'électricité [15]. En principe, plus l'information concernant ces risques est transmise rapidement, plus facilement ces mesures peuvent être mises en place.

2.2 Les modèles de prévision à base non physique

En général, un modèle statistique régressif permet de déduire une fonction d'approximation de nouvelles données à partir de données historiques. Dans les prochaines sous-sections, les différents modèles sont divisés en deux catégories, soient les régressions paramétriques et non paramétriques. Les modèles paramétriques reposent essentiellement sur l'estimation de paramètres tirés des données historiques, alors que les modèles non paramétriques sont basés directement sur les données. Cette séparation a été priorisée par rapport à la division entre les modèles purement statistiques et les modèles d'intelligence artificielle, dont font partie les modèles de forage de données, puisque la limite de ces deux domaines est encore aujourd'hui imprécise [72].

2.2.1 Les modèles paramétriques

L'utilisation des modèles stochastiques paramétriques dans la prévision de phénomènes hydrologiques remonte à plus d'un demi-siècle. Les travaux de Box et Jenkins tels que *Time Series Analysis: Forecasting and Control* [6] sur l'analyse et la prévision des séries chronologiques ont fortement influencé l'approche stochastique en hydrologie. Les modèles autorégressifs (AR), de moyennes mobiles (MA) et ceux découlant de leur combinaison (ARMA) qui y sont proposés sont, aujourd'hui encore, très populaires dans le domaine [1, 48, 60]. Toutefois, ces types de modèles présentent certaines lacunes pour les phénomènes hydrologiques de durée multi-annuelle. D'abord, les modèles de Box-Jenkins sont essentiellement des processus à courte mémoire ou *Short Range Dependence* (SRD), qui ne

tiennent compte que d'une fraction des occurrences passées. Par processus à courte mémoire, on entend une structure d'autocorrélation de ces modèles qui décroît de façon exponentielle à chaque décalage d'un pas de temps [30]. À titre d'exemple, cette structure peut être présente dans des chroniques hydrologiques au pas de temps annuel, comme le débit annuel moyen d'un bassin versant. En revanche, la plupart des processus hydrologiques présentent une persistance à long terme ou *Long Range Dependence* (LRD) c'est-à-dire une décroissance beaucoup plus lente de l'autocorrélation vers zéro à mesure que le décalage augmente [61, 67]. Cette structure d'autocorrélation peut être observée par la tendance des années humides et sèches à se regrouper en périodes humides et vice-versa [29].

Ensuite, les modèles autorégressifs et ses dérivés reposent en grande partie sur une hypothèse de normalité, alors que le type de distribution et le type d'autocorrélation représentant les processus hydrologiques évoluent en cours d'année à cause de leur comportement saisonnier. Les périodes de transition, causées par la saisonnalité, sont extrêmement difficiles à traiter par des techniques de dessaisonalisation généralement utilisées dans les modèles de type Box-Jenkins [35].

2.3 Exemple de modèles non paramétriques

Les réseaux de neurones présentent un exemple de modèle non paramétrique qui connaît un important regain d'intérêt surtout dans les domaines de l'intelligence artificielle et du forage de données. La modélisation par réseau de neurones artificiels est une technique axée sur

les données qui a retenu l'attention au cours des dernières années [62]. Dans de nombreux domaines, les réseaux de neurones se sont révélés efficaces pour simuler des systèmes complexes et identifier les relations non linéaires entre les intrants et des extrants sans pour autant chercher à en expliquer la nature [3]. Les premiers tests utilisant des réseaux de neurones dans le domaine de l'hydrologie ont été effectués au début des années 90 [28, 32, 43]. Ces premiers résultats prometteurs ont créé un certain engouement autour de cette méthode. Dans les dernières années, ce type de modèle a démontré à maintes reprises sa capacité à produire d'excellents résultats dans le domaine de l'hydrologie [53] et de l'hydro-électricité [11, 63]. Les plus récentes études sur ces modèles ont montré que ceux-ci sont rapides à développer et à utiliser pour la reconnaissance des patrons, la généralisation et la prévision des tendances.

Les réseaux de neurones possèdent plusieurs avantages sur les modèles stochastiques plus classiques et plusieurs études comparatives en ont démontré leur supériorité [36, 41, 64]. Bien que les modèles de réseaux de neurones semblent permettre de prédire les séries chronologiques dans le domaine de l'hydrologie et de l'hydro-électricité, ces derniers comportent aussi quelques inconvénients. Entre autres, tel que mentionné précédemment, la relation entre l'entrée et la sortie du modèle n'est pas expliquée, ce qui rend le modèle inapproprié pour diverses applications de scénarios. Aussi, la nature même de ces modèles repose sur une méthode de calibration de type essais et erreurs qui peut causer des enjeux liés à la lenteur de la vitesse d'apprentissage, au sur-apprentissage et à la forte tendance de ces modèles à atteindre et rester prisonnier d'un optimum local [10].

2.4 Modèles simplifiés

Une alternative stochastique et non paramétrique, couramment utilisée dans le domaine de l'hydrologie, est la méthode des *K* plus proches voisins (knn). Cette méthode se base sur la théorie de reconnaissance de schème qui remonte aux années 1950 et ne nécessite aucune hypothèse sur le type de dépendance à long ou court terme ni l'estimation de nombreux paramètres [56]. Cette méthode d'apprentissage par analogie est très populaire pour la reconnaissance de patrons et les prévisions de séries temporelles, son utilisation pour modéliser les processus hydrologiques d'écoulement en rivière a été bien documentée [37, 58, 74]. Par exemple, en hydrologie, pour une prévision de débit, l'hypothèse de base pour l'utilisation d'un modèle knn est que les séquences de débits historiques similaires à celles du passé récent fourniront des informations utilisables sur les débits qui se produiront dans un avenir rapproché. Le débit d'une rivière peut être perçu comme une agrégation de plusieurs processus météorologiques et hydrologiques tels que les précipitations sur le bassin versant, les pertes par évaporation et les variations d'humidité du sol. Sur une succession de mois, les précipitations et l'évaporation potentielle dépendront de l'état et de l'évolution des schémas de circulation atmosphérique à grande échelle. La réponse du débit de la rivière dépendra également de l'état et des caractéristiques du bassin versant [59]. Si aucune pluie ne tombe, le taux de diminution du débit peut être raisonnablement connu.

L'hypothèse sous-jacente à l'utilisation d'analogues dans un modèle knn comme méthode de prévision est qu'il existe des patrons ou trajectoires

particulières que le système hydro-climatique pourrait suivre et qui devraient se répéter. Ce type de modèle a été largement appliqué au domaine climatique [68], mais l'utilisation d'un seul analogue s'est révélé avoir une applicabilité limitée. En hydrologie, l'utilisation de plusieurs séquences historiques de débit entrant a été testée, avec des résultats similaires aux modèles plus complexes, sur des réservoirs et en rivières pour produire une prévision probabiliste des débits futurs [45].

2.5 Sélectionner les séries analogues

La méthode knn pose un problème quant à la définition des plus proches voisins. Contrairement aux données non chronologiques pour lesquelles une définition simple de la distance peut être utilisée directement, le choix de la mesure de similitude appropriée entre deux séries temporelles est un enjeu à considérer. La distance entre les séries chronologiques doit être soigneusement définie afin de refléter les similitudes de forme, d'amplitude et de synchronisme. Les différentes mesures de distance peuvent se diviser en deux catégories principales, selon qu'elles se basent sur la forme des séries ou leurs caractéristiques [18].

2.5.1 Mesure de similarité basée sur les caractéristiques des séries

Le premier type de mesure consiste à décomposer les séries temporelles en un vecteur de caractéristiques. Les propriétés des séries temporelles sont ensuite comparées pour en déduire la similarité [73]. Cette méthode permet de retirer la variable temporelle du problème et de comparer les séries de façon statique [71]. Un exemple très simple de cette méthode consiste à

représenter différentes séries temporelles selon leur moyenne et leur écart-type. Ainsi, peu importe la longueur initiale de la série, seulement ces deux valeurs sont comparées afin de trouver les plus proches voisins, réduisant la complexité du problème.

De nombreuses méthodes ont été développées pour aider à la sélection des caractéristiques [73]. Tout en limitant le nombre de valeurs à comparer, le vecteur de caractéristiques doit être en mesure de faciliter la bonne classification des séries temporelles pour le problème donné [20]. Les méthodes les plus populaires incluent le Lasso [65] et l'élimination récursive de caractéristiques [25]. Un choix judicieux de caractéristiques présente de nombreux avantages potentiels additionnels. En plus de réduire la dimensionnalité du problème, la sélection de caractéristiques peut permettre d'améliorer les performances de prévision, de faciliter la visualisation et la compréhension des données, de réduire le temps de calcul et les besoins de stockage de données [24].

2.5.2 Mesure de similarité basée sur la forme

Le second type de mesures de distance compare la forme globale des séries temporelles en fonction de leurs valeurs réelles. Deux sous-catégories peuvent être identifiées: les mesures en parallèle et les mesures élastiques. Les mesures en parallèle exigent que les séries chronologiques aient la même longueur et comparent, un pour un, chaque point temporel d'une série avec le point équivalent de la seconde série temporelle. Les mesures élastiques, quant à elles, sont plus souples et permettent de comparer des valeurs de plusieurs points temporels à plusieurs points temporels [70]. La mesure de similarité la plus courante dans l'étude de séries chronologiques

est probablement la distance euclidienne. La similitude entre deux séries données représente la somme des différences de chaque paire d'éléments correspondants. La distance euclidienne se démarque des autres mesures de similarité par sa faible complexité et son temps de calcul court [50]. Par contre, le manque de robustesse de cette méthode est un inconvénient majeur. Les séquences doivent obligatoirement avoir le même nombre de points. De plus, même si deux séries semblent similaires, la mesure est très sensible aux distorsions, décalages et différences d'amplitude [8]. Bien que l'incidence de ce dernier point puisse être atténuée par la normalisation des différents jeux de données [22], la distance euclidienne n'est pas nécessairement en mesure de détecter la similitude entre deux signaux subjectivement semblables.

Une autre mesure répandue, représentant la corrélation entre les séries temporelles, est basée sur le coefficient de Pearson. Cette mesure prend en compte les variations relatives autour de la moyenne dans le temps. Bien que contrairement à la distance euclidienne, le coefficient de Pearson soit basé sur la ressemblance de la forme générale entre deux séries, cette mesure est tout de même sensible au décalage temporel.

Afin de pallier aux différents problèmes des mesures de similarité point par point, plusieurs méthodes dites élastiques ont été développées. Ces mesures de distance ont l'avantage d'être robustes par rapport au décalage temporel et aux phénomènes d'accélération et de ralentissement observés dans les séries temporelles [8]. Une des mesures de déformation les plus simples, appelées la mesure d'édition, a été développée en déterminant le minimum d'opérations nécessaires pour rendre deux séries chronologiques

identiques [18]. Les opérations permises sont l'insertion, la suppression et la substitution [40]. Cette mesure est toutefois généralement utilisée pour déterminer la similarité entre deux mots en analysant leur série de caractères respectifs. Dans le cas des séries temporelles, utilisant des valeurs réelles, l'application de cette mesure est limitée puisqu'elle n'est pas sensible à la proximité des valeurs même. Une autre méthode élastique, appelée déformation temporelle dynamique (DTW), développée spécifiquement pour les séries numériques, offre une grande robustesse en remplaçant les comparaisons point par point par des comparaisons multiples entre plusieurs points des deux séries simultanément. Avec le DTW, les séries sont déformées de façon non linéaire pour trouver leur alignement optimal [54]. Cet aspect est primordial pour la recherche d'analogues dans les domaines de l'hydrologie et incidemment de la production hydro-électrique. En effet, quoique similaires, les phénomènes hydrologiques à long terme n'ont jamais exactement la même forme ni la même durée. Par exemple, sur une rivière donnée, la crue printanière se veut généralement cohérente d'année en année, guidée par les différents événements de fonte de neige, de variation de température et de précipitations. Cependant, le volume exact d'eau évacuée pendant la crue et la date à laquelle celle-ci débute sont variables d'année en année.

2.6 Remplissage de brèches

La visualisation de données permet de s'assurer que les séries chronologiques sont complètes et cohérentes avant de les utiliser dans un modèle statistique. Une attention particulière doit être portée à la sélection de la technique de remplissage de brèches pour combler des données

manquantes, celle-ci pourrait influencer les résultats finaux d'une étude statistique [26]. L'interpolation linéaire est parmi les techniques largement utilisées en hydrologie. Une équation est créée pour exprimer la variable dépendante en fonction de la variable indépendante, ici la production hydro-électrique et le temps. L'équation est ensuite appliquée sur les pas de temps où les données de la variable dépendante sont manquantes pour en estimer la valeur. Cette méthode diminue par contre la variabilité de la série [38].

La régression multivariée de son côté propose d'utiliser des données externes afin de combler les trous dans les séries chronologiques. Des relations sont donc trouvées entre la série à l'étude et d'autres séries qui lui sont corrélées [46], par exemple, la production hydro-électrique d'une centrale, la mesure des débits dans le bief amont et la hauteur de chute. Quoiqu'elle permette de conserver une certaine variabilité, cette technique implique l'utilisation de données exogènes, autres que celles de production hydro-électrique.

La pondération inverse à la distance, *inverse distance weigthing,* (IDW) est un autre exemple de méthode qui utilise les données provenant de sources géographiquement proches pour remplir les brèches des séries. Cette méthode requiert un bon maillage spatial des sources de données. Sinon, les prévisions risquent d'être surestimées [26]. Par exemple, dans le cas de séries de données provenant de station de mesures de pluie, si la station sur laquelle l'IDW est appliquée est entourée de cinq stations relativement rapprochées, chaque fois qu'une de ces stations montre des précipitations, même si elle est la seule des cinq, la station sur laquelle l'IDW est appliquée montrera des précipitations non nulles.

3 MÉTHODOLOGIE

Ce chapitre définit les principales étapes pressenties, menant à la création d'un modèle statistique de type knn permettant de faire des prévisions de la production hydro-électrique basées sur les séries analogues. Les horizons prévisionnels testés sont de 3 et 6 mois. Dans ce document, X et Y indiquent, respectivement, une série temporelle explicative et une série temporelle d'intérêt. Par exemple, Y représente l'année en cours et X représente une année similaire à Y. Ces deux séries, qui n'ont pas nécessairement la même longueur, sont constituées d'éléments $[x_1, x_2, \dots, x_t, x_{t+1}, \dots, x_{n-1}, x_n]$ ou $[y_1, y_2, \dots, y_t, y_{t+1}, \dots, y_{m-1}, y_m]$ qui représentent une valeur numérique au pas de temps t.

Les données sont d'abord prétraitées afin d'en réduire le bruit. Ensuite, les séries explicatives sont comparées à la série d'intérêt pour mesurer leur similarité. Une régression multivariée est calculée à l'aide des séries explicatives les plus semblables et, subséquemment, des probabilités d'occurrence sont tirées à l'aide de l'échantillon restreint de séries explicatives similaires à la série d'intérêt.

Les différents paramètres du modèle multivarié sont par la suite optimisés par simulation rétrospective aussi appelée «*hindcasting*». À cette étape, toutes les séries explicatives sont tour à tour utilisées pour tester le modèle et valider ses résultats de prévision par rapport aux valeurs réelles. L'ajustement des paramètres est guidé par l'amélioration observée de certaines mesures de performance.

3.1 Prétraitement

Afin de faciliter la procédure de reconnaissance de schème définissant l'évolution d'une série temporelle, plusieurs transformations peuvent être appliquées à l'ensemble de données. Le lissage permet de réduire le bruit causé par les fluctuations rapides d'une série de données pour en faire ressortir les tendances générales. Une des méthodes employées est la moyenne mobile. Une moyenne mobile de fenêtre *f* est calculée en divisant par *f* la somme des *f* plus récentes données. Cette moyenne est recalculée à chaque période en supprimant les données les plus anciennes et en ajoutant les plus récentes, de sorte que la moyenne évolue avec la série [21]. Une moyenne mobile typique est décrite par l'équation **(1)** :

$$y_t = \frac{1}{f}\sum_{i=0}^{f-1} x_{f-i} \qquad \textbf{(1)}$$

L'opération renvoie un nouvel ensemble de données *y*, où chaque pas de temps *t* représente la moyenne d'un sous-ensemble consécutif de taille *n* de la série temporelle d'origine *x*.

3.2 Recherche de similarité

Les séries hydrologiques sont souvent considérées comme auto-corrélées et saisonnières, les patrons sont généralement semblables inter-annuellement, mais varient grandement intra-annuellement [51]. Afin d'extraire un maximum d'information provenant de l'ensemble des données historiques, une méthode de reconnaissance de schèmes, ou un patron qui

décrit l'évolution dans le temps de la production hydro-électrique sur une année, est nécessaire. Pour ce faire, les différentes années d'une même série de données peuvent être considérées comme des variables supplémentaires permettant d'établir des prévisions [27].

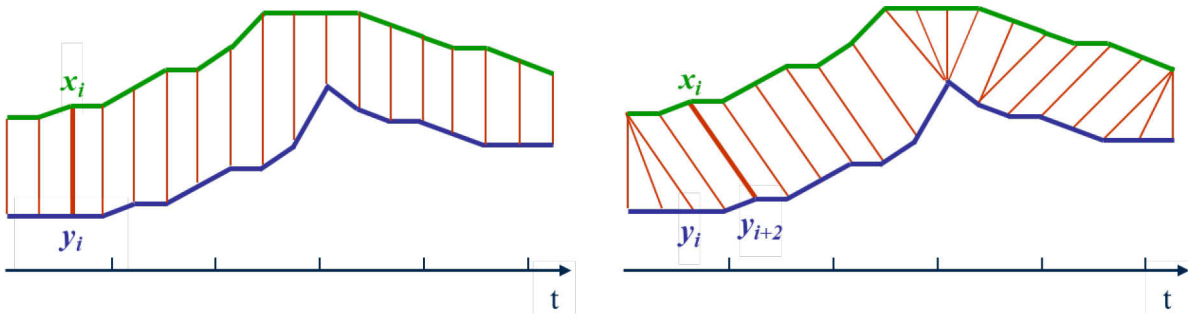
Une fois divisées en plusieurs variables annuelles, les séries peuvent être modifiées afin de mieux les comprendre et d'en extraire plus facilement les schèmes. Ces modifications permettent d'en faire ressortir les similitudes. La déformation temporelle dynamique, «*dynamic time warping*», (DTW) est une méthode non linéaire développée pour la reconnaissance vocale, mais qui est maintenant répandue dans bien des domaines, dont l'hydrologie [54]. La DTW est utilisée pour déformer et déplacer deux séries temporelles afin d'améliorer et ultimement de quantifier leur similitude, au lieu de mesurer la distance entre chacun des points correspondants des séries. Cette mesure est appelée la distance euclidienne d, comme le montre l'équation (2) :

$$d = |x_i - y_i| \qquad \textbf{(2)}$$

où x_i et y_i sont des réalisations de séries temporelles différentes au pas de temps *i*. Dans la présente étude, x_i et y_i sont différentes années de séries de données de production quotidienne du même système.

La production d'hydro-électricité est principalement basée sur les événements hydrologiques (fonte des neiges, sécheresse, tempête, etc.). Étant donné que l'amplitude et la durée de ces événements peuvent varier, l'utilisation d'une méthode de comparaison linéaire, comme la distance

euclidienne, pourrait manquer des patrons similaires [44]. La figure 3.1 montre la différence de méthodologie entre la distance euclidienne et une mesure élastique comme le DTW.



a) b)

Figure 3.1 Comparaison de différentes mesures de similarité a) distance euclidienne, b) une mesure élastique [66]

La distance mesurée de façon élastique permet de trouver un alignement plus intuitif entre deux séries même si elles sont décalées de quelques pas de temps.

L'application du DTW s'effectue en trois étapes. D'abord, une matrice est formée pour calculer la distance euclidienne entre chaque combinaison possible de valeurs des séries temporelles *X* et *Y* [44], comme indiqué dans l'équation (3) et la représentation de la matrice des distances euclidiennes (4) :

$$Y = [y_1 \quad y_2 \quad \cdots \quad y_m];$$

$$X = [x_1 \quad x_2 \quad \dots \quad x_n]$$

$$d_{ij} = |x_i - y_j| \qquad \textbf{(3)}$$

$$\begin{bmatrix} d_{1m} & d_{2m} & \dots & d_{nm} \\ \dots & \dots & \dots & \dots \\ d_{12} & d_{22} & \dots & d_{n2} \\ d_{11} & d_{21} & \dots & d_{n1} \end{bmatrix} \qquad \textbf{(4)}$$

où *n* représente la longueur des séries étudiées. Dans le cas de séries annuelles journalières, *n* serait égal à 365, ou moins en cas de données manquantes; Y représente la variable d'intérêt; y_1 à y_n sont les réalisations de la variable d'intérêt; X représente une variable explicative; x_1 à x_n sont les réalisations de la variable explicative. Chaque point d_{ij} de la matrice (4) correspond à la distance euclidienne entre les points x_i et y_j.

Ensuite, à l'aide de la programmation dynamique, une matrice de distance cumulée *D* de mêmes dimensions que la matrice (4) est construite en utilisant l'équation (5) :

$$\forall_{ij}\ D_{ij} = d_{ij} + min \begin{cases} \text{i: } D_{i-1,j-1} \\ \text{ii: } D_{i-1,j} \\ \text{iii: } D_{i,j+1} \end{cases} \qquad \textbf{(5)}$$

$$\begin{bmatrix} D_{1m} & D_{2m} & \dots & D_{nm} \\ \dots & \dots & \dots & \dots \\ D_{12} & D_{22} & \dots & D_{n2} \\ D_{11} & D_{21} & \dots & D_{n1} \end{bmatrix} \qquad \textbf{(6)}$$

De cette façon, chaque point D_{ij} correspond à la distance cumulée la plus courte pour atteindre la position *i, j* de la matrice (6). Ce qui signifie que la mesure de similarité entre les séries *X* et *Y* est donnée par la valeur de D_{nm}. Toutefois, l'algorithme du DTW doit être supervisé en bloquant certaines zones de la matrice de distance euclidienne pour éviter des résultats trompeurs découlant de corrélations irréalistes. Par exemple, pour des séries temporelles de débit en rivière, en laissant une trop grande liberté à l'algorithme, celui-ci pourrait tenter de corréler une crue d'automne causée par un épisode de pluie intense à une crue printanière causée par la fonte des neiges. Pour éviter ce problème, les distances cumulées D_{ij} de la matrice (6) doivent respecter au moins ces trois règles [57] :

i Être monotone, les séries *X* et *Y* ne peuvent pas reculer dans le temps;

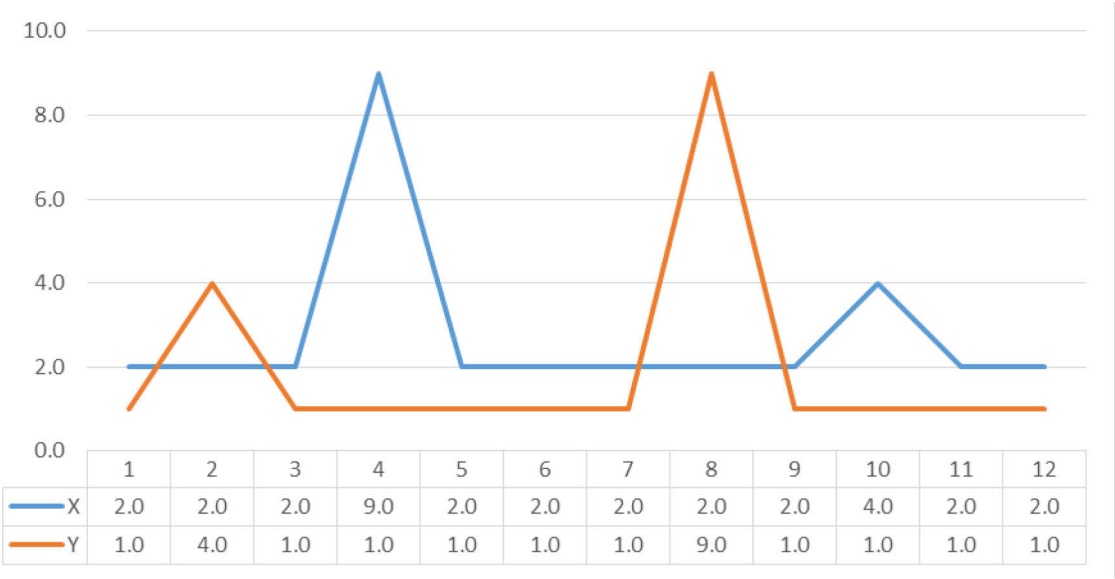
ii Croître de façon unitaire, les séries *X* et *Y* ne peuvent avancer de plus d'un pas de temps à la fois;

iii Localisées, le trajet déformé doit demeurer près de la diagonale.

Les deux premières règles sont automatiquement suivies par l'application de l'équation (8). Une des façons les plus simples de circonscrire la DTW et de s'assurer du respect de la troisième règle est d'imposer une fenêtre de déformation *w*. Ainsi, la déformation des séries temporelles ne peut dépasser quelques pas de temps [14].

3.2.1 Exemple d'application de l'algorithme DTW

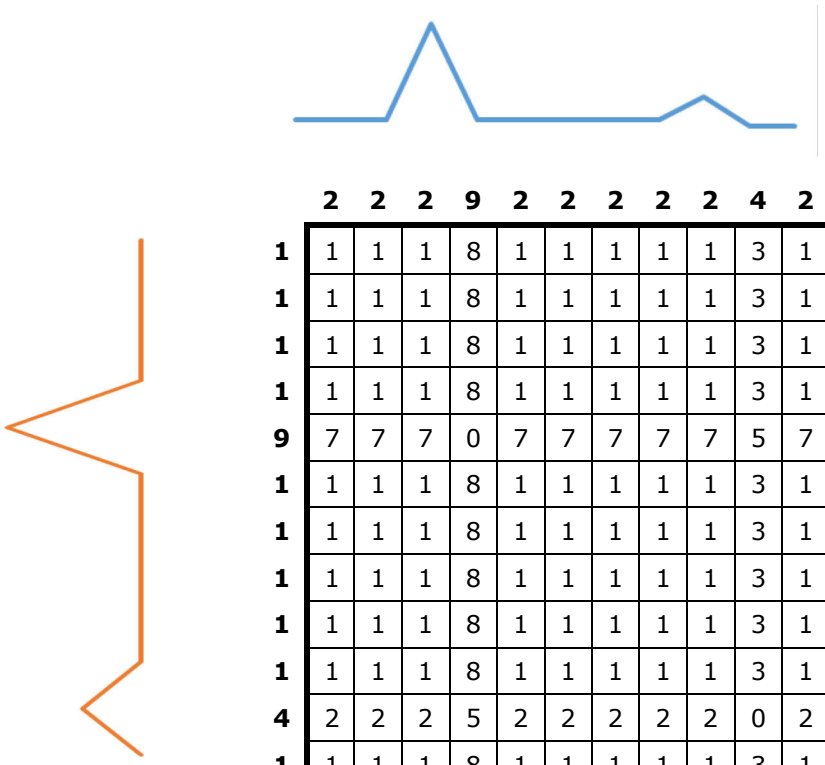
En supposant deux séries fictives *X* et *Y*, dont les valeurs à chaque pas de temps et leur représentation graphique sont définies à la figure 3.2.



	1	2	3	4	5	6	7	8	9	10	11	12
X	2.0	2.0	2.0	9.0	2.0	2.0	2.0	2.0	2.0	4.0	2.0	2.0
Y	1.0	4.0	1.0	1.0	1.0	1.0	1.0	9.0	1.0	1.0	1.0	1.0

Figure 3.2 Séries X et Y

Ces deux séries présentent plusieurs similitudes, chacune montre un pic en début et en fin de séquence, mais leurs pointes sont de différentes amplitudes et ne sont pas alignées. Dans un cas semblable, le DTW, contrairement à la distance euclidienne, devrait permettre de capter ces similitudes. Pour ce faire, une matrice de distance est construite à l'aide de l'équation (3).



	2	**2**	**2**	**9**	**2**	**2**	**2**	**2**	**2**	**4**	**2**	**2**
1	1	1	1	8	1	1	1	1	1	3	1	1
1	1	1	1	8	1	1	1	1	1	3	1	1
1	1	1	1	8	1	1	1	1	1	3	1	1
1	1	1	1	8	1	1	1	1	1	3	1	1
9	7	7	7	0	7	7	7	7	7	5	7	7
1	1	1	1	8	1	1	1	1	1	3	1	1
1	1	1	1	8	1	1	1	1	1	3	1	1
1	1	1	1	8	1	1	1	1	1	3	1	1
1	1	1	1	8	1	1	1	1	1	3	1	1
1	1	1	1	8	1	1	1	1	1	3	1	1
4	2	2	2	5	2	2	2	2	2	0	2	2
1	1	1	1	8	1	1	1	1	1	3	1	1

Figure 3.3 Matrice de distance point par point entre les séries X et Y

Les traits orange et bleu sont montrés à titre indicatif seulement. Les nombres en gras représentent les valeurs à chaque pas de temps des séries, *X* pour la ligne d'en-tête et *Y* pour la colonne d'en-tête. L'échelle temporelle est positive vers le haut et vers la droite. Chaque élément d_{ij} à l'intérieur de la matrice montre la différence entre les points associés x_i et y_j. Ensuite, la mise en place de la matrice des distances cumulées consiste en 4 étapes.

1. Déduire la valeur du premier élément $D_{1,1}$ situé dans le coin inférieur gauche à l'aide de l'équation (7) :

$$D_{1,1} = d_{1,1} \quad \textbf{(7)}$$

2. Calculer la première ligne en bas de la matrice en utilisant l'équation (8) :

$$D_{i,1} = d_{i,1} + D_{i-1,1} \quad \textbf{(8)}$$

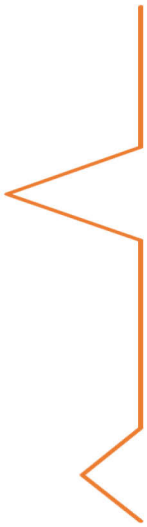
3. Calculer la première colonne en utilisant l'équation (9) :

$$D_{1,j} = d_{1,j} + D_{1,j-1} \quad \textbf{(9)}$$

4. Chiffrer tous les autres éléments à l'aide de l'équation (5) :

$$\forall_{ij}\ D_{ij} = d_{ij} + min \begin{cases} \text{i: } D_{i-1,j-1} \\ \text{ii: } D_{i-1,j} \\ \text{iii: } D_{i,j+1} \end{cases} \quad \textbf{(5)}$$

Le résultat final est montré à la figure 3.4.



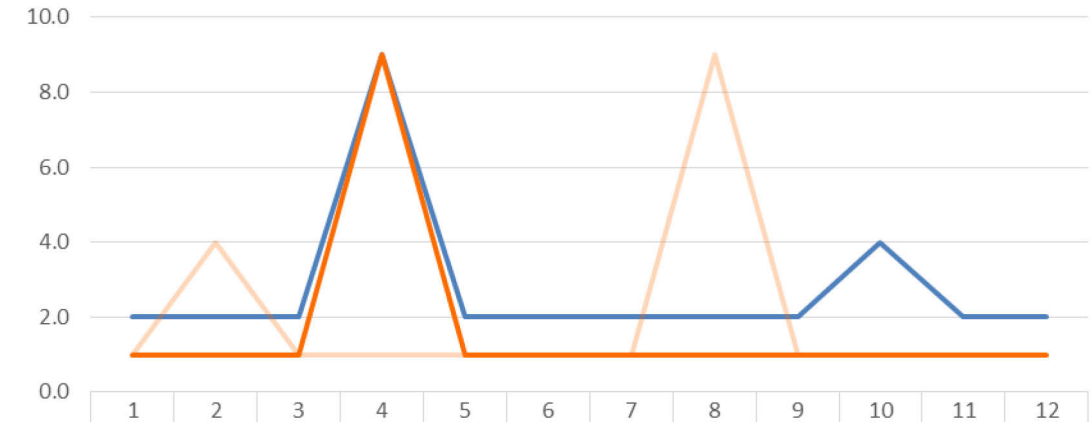
	2	2	2	9	2	2	2	2	2	4	2	2
1	19	19	19	26	12	12	12	12	13	16	17	18
1	18	18	18	25	11	11	11	12	13	16	17	18
1	17	17	17	24	10	10	11	12	13	16	17	18
1	16	16	16	16	9	10	11	12	13	16	17	18
9	15	15	15	8	15	20	20	20	20	18	23	24
1	8	8	8	15	13	13	13	13	13	16	17	18
1	7	7	7	14	12	12	12	12	13	16	17	18
1	6	6	6	13	11	11	11	12	13	16	17	18
1	5	5	5	12	10	10	11	12	13	16	17	18
1	4	4	4	12	9	10	11	12	13	16	17	18
4	3	3	4	8	10	12	14	16	17	16	18	20
1	1	2	3	11	12	13	14	15	16	19	20	21

Figure 3.4 Matrice de distance cumulée

Les en-têtes en gras représentent toujours les valeurs originales à chaque pas de temps des séries. Dans cette matrice, chaque élément D_{ij} présente la distance cumulative la plus courte pour atteindre ce point. Les cases colorées montrent la trajectoire de déformation. Cette trajectoire permet de constater l'étendue des transformations qui ont été nécessaires pour obtenir un alignement optimal des deux séquences. À chaque pas de temps, la position suivante est donnée par l'une des trois options suivantes:

1. Si la case suivante de la ligne colorée est en diagonale, y avance normalement : $(y_1, y_2, \dots, y_t, y_{i+1}, y_{i+2}, \dots)$
2. Si la case suivante est à la verticale, y saute un pas de temps : $(y, y_2, \dots, y_i, y_{i+2}, y_{i+3}, \dots)$
3. Si la case suivante est à l'horizontale, y se répète : $(y, y_2, \dots, y_i, y_i, y_{i+1}, \dots)$

L'alignement optimal trouvé par le DTW est montré à la figure 3.5.

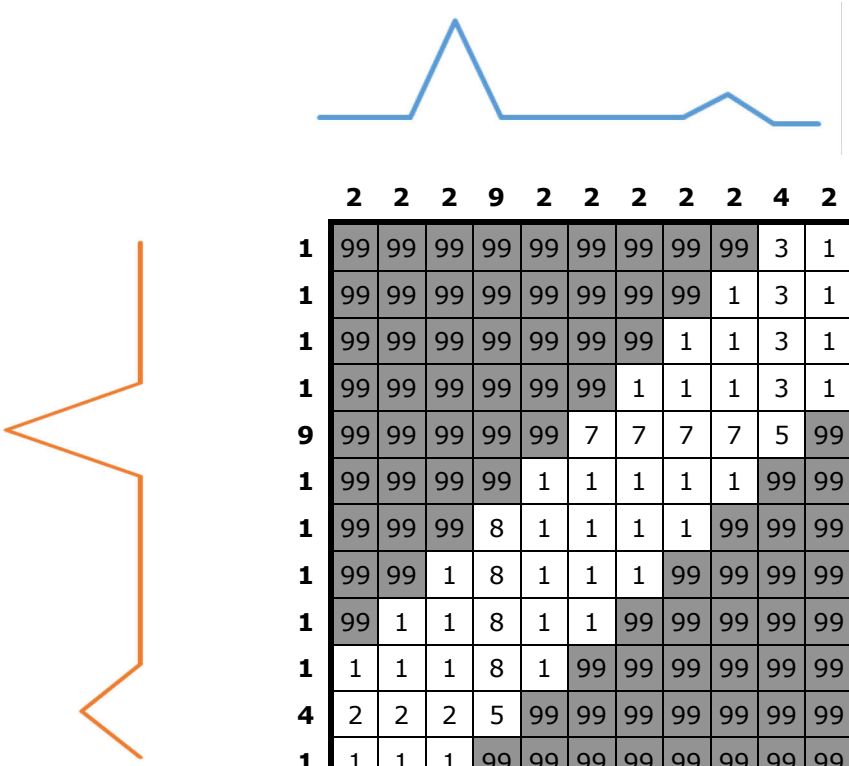


	1	2	3	4	5	6	7	8	9	10	11	12
X	2.0	2.0	2.0	9.0	2.0	2.0	2.0	2.0	2.0	4.0	2.0	2.0
Y	1.0	4.0	1.0	1.0	1.0	1.0	1.0	9.0	1.0	1.0	1.0	1.0
Y'	1.0	1.0	1.0	9.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Figure 3.5 Séries X, Y et Y' déformée à l'aide du DTW

X et *Y* sont les séries originales, la couleur de la série *Y* a été adoucie pour éviter la confusion, et *Y'* représente la séquence déformée. L'algorithme du DTW a trouvé que le meilleur alignement était de décaler complètement *Y* afin que les deux pics de mêmes amplitudes soient parfaitement superposés. Ce résultat est contre-intuitif, un alignement plus réaliste aurait été de transformer les séries afin que les pics de début et de fin de

séquences soient en phase pour les deux séries. D'où l'importance de la fenêtre de déformation *W* qui permet d'encadrer le DTW et de circonscrire la déformation à une zone définie. Une des façons les plus simples d'appliquer cette contrainte est de remplacer les valeurs de la matrice de distance point par point qui se trouvent à une distance plus grande que *W* de la diagonale par une valeur très élevée.

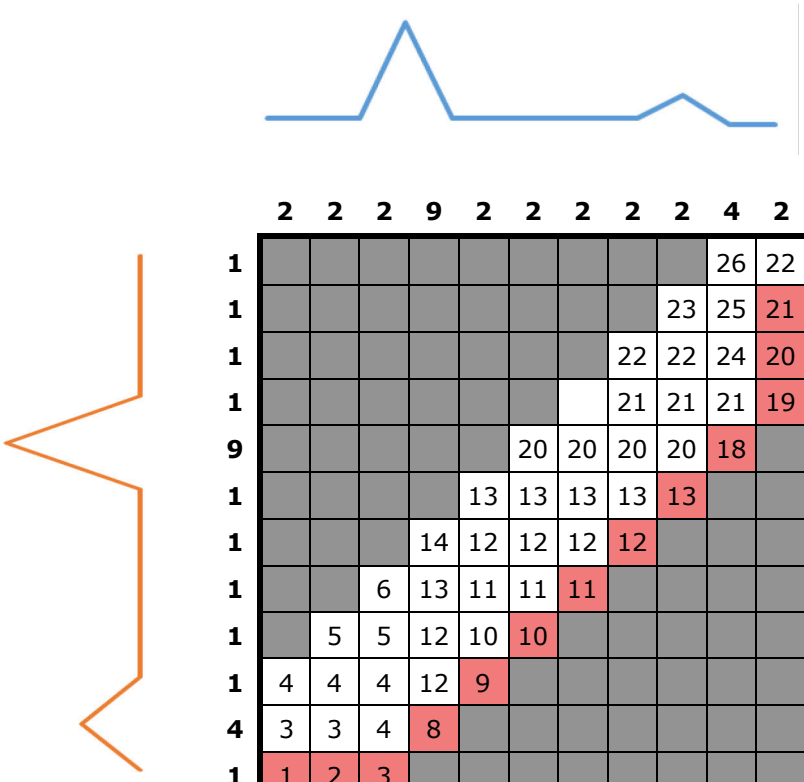


	2	**2**	**2**	**9**	**2**	**2**	**2**	**2**	**2**	**4**	**2**	**2**
1	99	99	99	99	99	99	99	99	99	3	1	1
1	99	99	99	99	99	99	99	99	1	3	1	1
1	99	99	99	99	99	99	99	1	1	3	1	1
1	99	99	99	99	99	99	1	1	1	3	1	99
9	99	99	99	99	99	7	7	7	7	5	99	99
1	99	99	99	99	1	1	1	1	1	99	99	99
1	99	99	99	8	1	1	1	1	99	99	99	99
1	99	99	1	8	1	1	1	99	99	99	99	99
1	99	1	1	8	1	1	99	99	99	99	99	99
1	1	1	1	8	1	99	99	99	99	99	99	99
4	2	2	2	5	99	99	99	99	99	99	99	99
1	1	1	1	99	99	99	99	99	99	99	99	99

Figure 3.6 Matrice de distance point par point entre les séries X et Y bornée par une fenêtre *W*

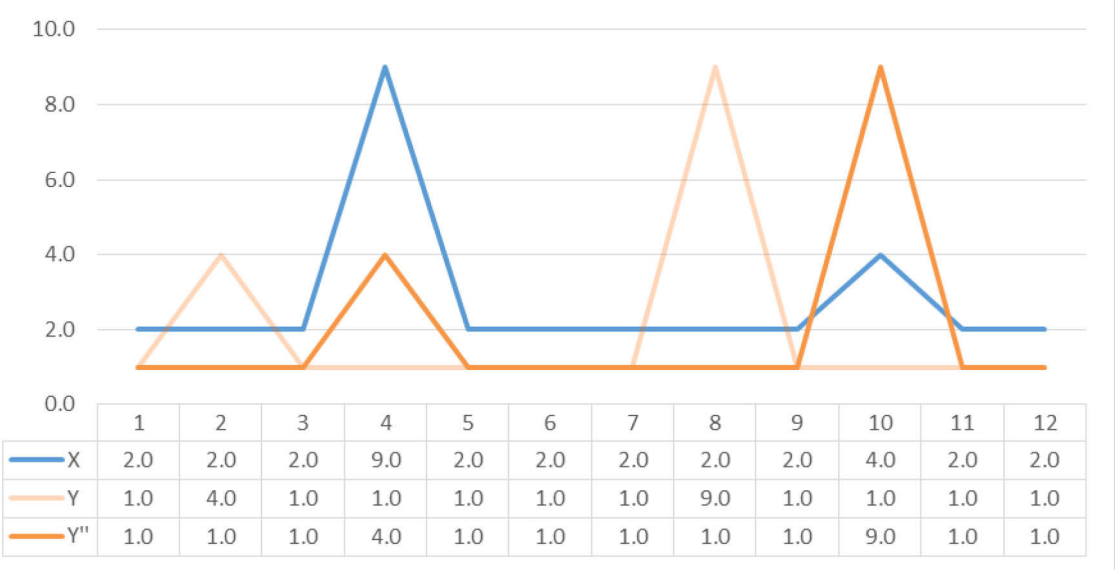
Dans le cas de la figure 3.6, la fenêtre choisie présente un paramètre *W* égal à 2. Ce paramètre est généralement choisi par essai erreur avec une

connaissance générale des données comparées. De cette façon la trajectoire de déformation ne peut pas dévier de façon significative de la diagonale, comme le montre la figure 3.7. Le nouvel alignement optimal est présenté à la figure 3.8.



	2	2	2	9	2	2	2	2	2	4	2	2
1										26	22	22
1									23	25	21	21
1								22	22	24	20	20
1								21	21	21	19	
9						20	20	20	20	18		
1					13	13	13	13	13			
1				14	12	12	12	12				
1			6	13	11	11	11					
1		5	5	12	10	10						
1	4	4	4	12	9							
4	3	3	4	8								
1	1	2	3									

Figure 3.7 Matrice de distance cumulée bornée par la fenêtre *W* = 2



	1	2	3	4	5	6	7	8	9	10	11	12
X	2.0	2.0	2.0	9.0	2.0	2.0	2.0	2.0	2.0	4.0	2.0	2.0
Y	1.0	4.0	1.0	1.0	1.0	1.0	1.0	9.0	1.0	1.0	1.0	1.0
Y''	1.0	1.0	1.0	4.0	1.0	1.0	1.0	1.0	1.0	9.0	1.0	1.0

Figure 3.8 Séries X, Y et Y'' déformée à l'aide du DTW borné

La trajectoire de déformation de la figure 3.7 diffère de celle montrée précédemment puisque les cases noircies sont en quelque sorte hors limite pour le DTW. La distance cumulée totale $D_{12,12}$ pour cet exemple, affichée dans le coin supérieur droit, est plus grande que celle de la figure 3.4. Ces différences indiquent que la fenêtre W a influencé les résultats de l'algorithme, comme le prouve la figure 3.8, où les pics de début et de fin de séquences sont bien alignés malgré leur différence d'amplitude.

3.3 Modèle de régression

Une fois les séries similaires à la série d'intérêt trouvées, un modèle de régression peut être utilisé pour définir le modèle statistique knn. Les futures occurrences de la série d'intérêt sont ainsi estimées selon les observations des séries explicatives. Les modèles de régression permettent d'effectuer des prévisions afin d'étendre différentes séries de données [2].

La régression multiple étudie la relation entre une variable d'intérêt et plusieurs variables explicatives. Les équations (11) à (13) montrent la relation entre la variable endogène et les variables explicatives dans une régression multivariée [12] :

$$y_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + \cdots + a_p x_{ip} \quad \textbf{(10)}$$

$$Y = y_1, y_{2,} \ldots, y_i, y_{i+1}, \ldots, y_n \quad \textbf{(11)}$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{pmatrix} \quad \textbf{(12)}$$

Où *n* représente la longueur des séries étudiées. Par exemple, dans le cas de série annuelle journalière, *n* serait égal à 365; *i* est un pas de temps spécifique qui peut aller de 1 jusqu'à n; *p* représente le nombre de variables explicatives; Y représente la variable endogène; y_1 à y_n sont les réalisations de la variable endogène; X_1 représente la première variable explicative; $x_{1,1}$ à $x_{n,1}$ sont les réalisations de la première variable explicative et ainsi de suite jusqu'à la *p* ième variable explicative X_p; a_0 à a_p indique le facteur de pondération associé à chaque variable explicative.

3.4 Post-Traitement

Les séries explicatives utilisées pour effectuer la prévision de la série d'intérêt peuvent être utilisées comme un ensemble de données sur lequel

il est possible d'effectuer des analyses statistiques. Ces analyses statistiques sont nécessaires afin d'analyser l'incertitude reliée aux résultats. À cet ensemble, à chaque pas de temps, l'ajustement d'une distribution est possible. L'application d'une distribution normale est largement utilisée dans le domaine de la statistique vu sa simplicité. Par contre, avec des données de production d'hydro-électricité, étant donné les limitations physiques d'une turbine, sa production minimale est bornée à 0 MW, alors que sa production maximale est limitée par sa capacité nominale. L'utilisation d'une distribution bornée telle que la distribution bêta qui permet de mieux contraindre la distribution de probabilité [34] devrait être préconisée. La distribution bêta est caractérisée par les facteurs α et β, influençant son asymétrie et son aplatissement. La figure 3.9 montre quelques exemples de l'effet des facteurs α et β ([α, β]) sur la forme de la distribution.

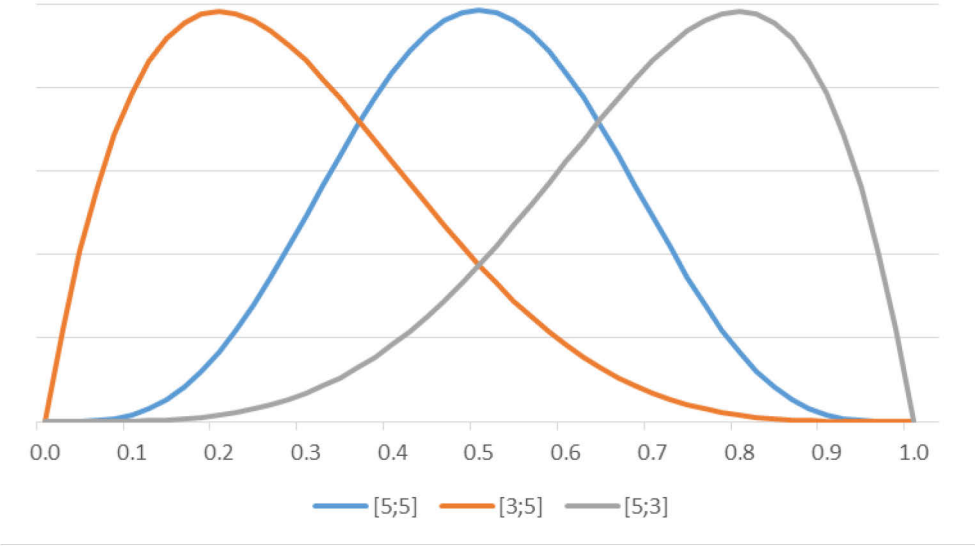


Figure 3.9 différentes formes de distribution bêta

Les termes α et β de la distribution bêta sont facilement calculés à partir des données statistiques simples comme la moyenne et l'écart type de la distribution tels que montrés par les équations (13) et (14) :

$$\mu = \frac{a\alpha + b\beta}{\alpha + \beta} \quad \textbf{(13)}$$

$$\sigma^2 = \frac{\alpha\beta(b-a)^2}{(\alpha+\beta)^2(1+\alpha+\beta)} \quad \textbf{(14)}$$

3.5 Optimisation de paramètres

Le modèle knn ainsi mis en place, voir les équations (10), (11) et (12) présente plusieurs paramètres à optimiser afin de maximiser sa capacité à produire des prévisions de qualité. Les trois principaux sont le nombre *k* de séries analogues à utiliser pour effectuer les prévisions; la longueur de séries ou, en d'autres mots, le nombre de pas de temps *T* à utiliser afin de trouver ses séries similaires à l'aide du DTW et la dimension de la fenêtre *W* qui limite la déformation autorisée au DTW pour maximiser l'alignement de ces séries.

Afin de trouver ces paramètres optimaux, le modèle est utilisé sur les données historiques. Les résultats sont ensuite comparés aux données réelles selon certaines mesures de performance graphique et numérique. Ces étapes sont répétées en boucle, en modifiant les paramètres de façon incrémentale à chaque itération, jusqu'à l'obtention d'un jeu de paramètres présentant les meilleurs résultats.

3.5.1 Simulation rétrospective

La simulation rétrospective aussi appelée «*hindcasting*» est une technique qui permet de valider un modèle en l'utilisant sur des données historiques. Les sorties de modèles sont ensuite comparées aux valeurs réellement observées.

3.5.2 Mesures de performance

Avec l'objectif de juger de façon objective la qualité des prévisions fournies par ce modèle, l'utilisation de deux types de mesures est prévue : la vérification graphique à l'aide d'histogrammes des transformations par fonction de répartition *Probability Integral Transform* (PIT) et la vérification numérique avec la racine carré de l'erreur quadratique moyenne (RMSE).

L'histogramme PIT permet d'évaluer la calibration d'un système de prévision probabiliste. D'autres mesures comme le score de probabilité continue classée (CRPS) ont été envisagées, mais le PIT a été sélectionné pour sa simplicité et son faible coût en temps de calcul. En se basant sur la prémisse que la répartition des résultats selon un intervalle régulier de probabilité d'occurrence devrait être uniforme [16], le PIT donne un indice concret et visuel de la sur ou sous dispersion et du biais possible des résultats d'un modèle. Cet histogramme est utilisé pour valider que la distribution bêta utilisée représente bien la réalité. Ainsi, pour une série d'intérêt *Y* et une série de résultats de modèle *Y'*, la distribution des occurrences y_1 à y_n devrait être similaire à celle des occurrences y'_1 à y'_n.

La figure 3.10 montre les différents cas de figure possibles des histogrammes PIT.

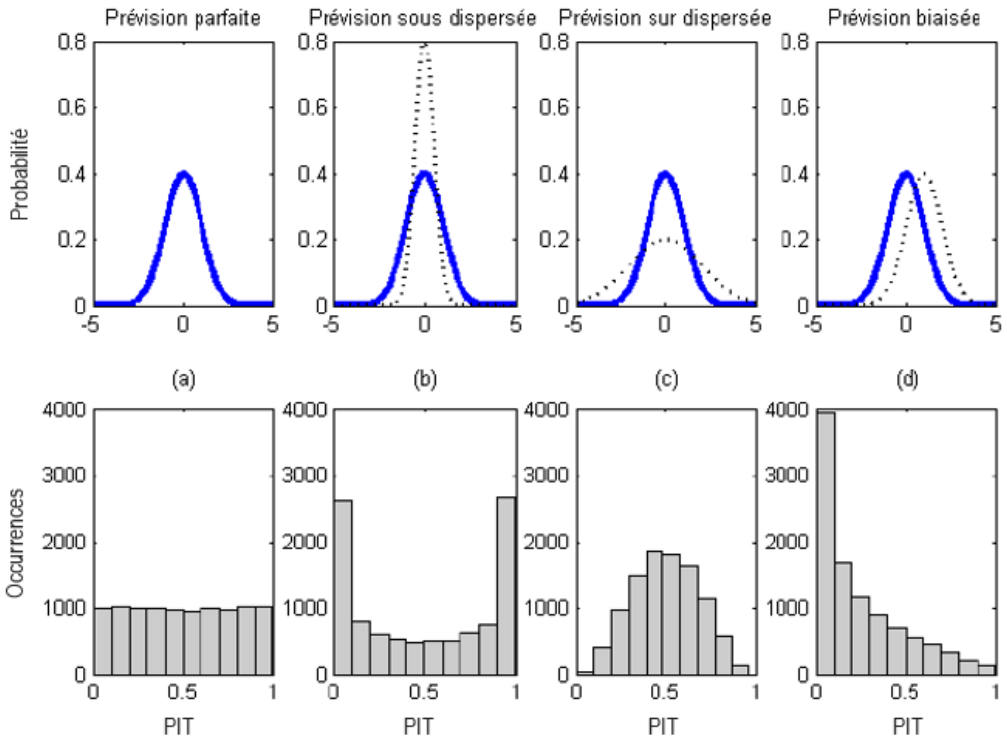


Figure 3.10 Quelques formes typiques de l'histogramme PIT : (a) prévision bien calibrée ; b) sous-dispersion ; (c) sur-dispersion; (d) biais [55]

À la Figure 3.10, sur les graphiques du haut, les traits bleus représentent une distribution normale des prévisions et le pointillé quelques exemples de distribution inexacte. Les histogrammes en gris montrent pour chaque exemple les occurrences par décile des prévisions. Sur une échelle T comprenant les pas de temps t à T chacune des valeurs réelles y_t à y_T sont

catégorisées selon la distribution. Chaque valeur est ainsi classée selon sa probabilité de dépassement estimée à l'aide de la distribution bêta. Ces quantiles sont alors comparés entre eux afin de s'assurer que le nombre d'occurrences qu'ils contiennent est uniforme.

La racine carrée de l'erreur quadratique moyenne (RMSE) représente l'écart type des erreurs de prévision. Ce qui permet d'indiquer la concentration des données autour de la droite de meilleur ajustement. La racine carrée de l'erreur quadratique moyenne est couramment utilisée pour vérifier les résultats expérimentaux de modèle de prévision et de régression. Le RMSE est calculé selon l'équation (15) :

$$RMSE = \sqrt{\sum_{i=t}^{T} \frac{(y_i - y'_i)^2}{i}} \qquad \textbf{(15)}$$

Les résultats de cette mesure sont exprimés avec les mêmes unités que la prévision. Une valeur faible de RMSE indique un meilleur ajustement.

4 ÉTUDE DE CAS

4.1 Données utilisées

Pour construire le modèle de prévision, les données de production hydro-électrique journalière de 181 centrales, réparties sur 68 systèmes hydriques au Canada et aux États-Unis ont été utilisées. La production de ces centrales a été agrégée par système hydrique dans le but d'en réduire le bruit et de permettre au modèle de suivre les tendances et trajectoires générales d'un bassin versant plutôt que les fluctuations d'une seule centrale. Letableau 4.1 liste ces rivières et la figure 4.1 montre l'étendue du territoire couvert par les centrales de ces systèmes hydriques.

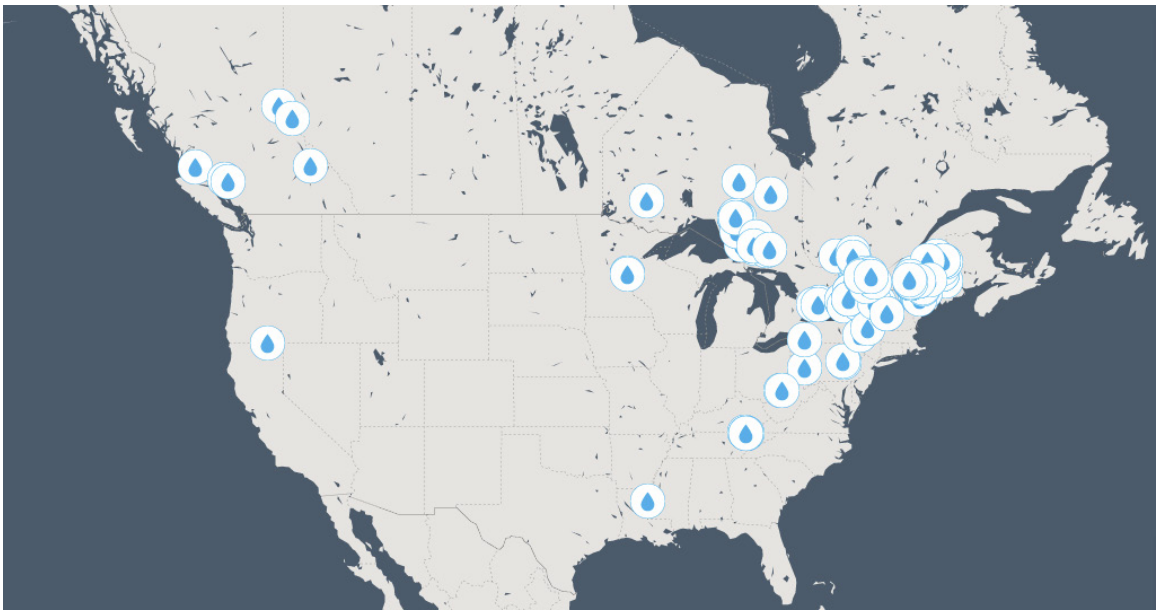


Figure 4.1 Répartitions spatiales des différentes centrales hydro-électriques [17]

Tableau 4.1 Systèmes hydriques utilisés pour le modèle de prévisions [17]

Pays	Région	Rivière	Capacité installée (MW)
États-Unis	Californie	Pit	30
États-Unis	**Californie**	**Total**	**30**
États-Unis	Midwest et sud-est	Clarion	53
États-Unis	Midwest et sud-est	Kanawha	6
États-Unis	Midwest et sud-est	Little Tennessee	376
États-Unis	Midwest et sud-est	Mississippi	192
États-Unis	Midwest et sud-est	Susquehanna	60
États-Unis	Midwest et sud-est	Upper Mississippi	28
États-Unis	**Midwest et sud-est**	**Total**	**1175**
États-Unis	Nouvelle Angleterre	Androscoggin	186
États-Unis	Nouvelle Angleterre	Deerfield	610
États-Unis	Nouvelle Angleterre	Kennebec	227
États-Unis	Nouvelle Angleterre	Penobscot	174
États-Unis	Nouvelle Angleterre	Saco	61
États-Unis	Nouvelle Angleterre	Union	9
États-Unis	**Nouvelle Angleterre**	**Total**	**1273**
États-Unis	New York	Beaver	48
États-Unis	New York	Black	40
États-Unis	New York	East Canada Creek	25
États-Unis	New York	Fish Creek	2
États-Unis	New York	Hoosic	18
États-Unis	New York	Hudson	111
États-Unis	New York	Mohawk	38
États-Unis	New York	NYS Barge Canal	3
États-Unis	New York	Oak Orchard Creek	6
États-Unis	New York	Oswegatchie	34
États-Unis	New York	Oswego	32
États-Unis	New York	Raquette	180
États-Unis	New York	Sacandaga	61
États-Unis	New York	Salmon North	5
États-Unis	New York	Salmon South	39
États-Unis	New York	Saranac	2
États-Unis	New York	Seneca	1
États-Unis	New York	St-Regis	7
États-Unis	**New York**	**Total**	**656**
États-Unis	**Total**	**Total**	**3134**

Pays	Région	Rivière	Capacité installée (MW)
Canada	Colombie Britannique	East Twin Creek	2
Canada	Colombie Britannique	Hystad Creek	6
Canada	Colombie Britannique	Kokish	45
Canada	Colombie Britannique	Lois	37
Canada	Colombie Britannique	Powell	38
Canada	**Colombie Britannique**	**Total**	**173**
Canada	Ontario	Aux Sables	4
Canada	Ontario	Magpie	43
Canada	Ontario	Michipicoten	104
Canada	Ontario	Mississagi	488
Canada	Ontario	Montreal	150
Canada	Ontario	Nagagami	19
Canada	Ontario	Seine	10
Canada	Ontario	Serpent	7
Canada	Ontario	St-Mary's	52
Canada	**Ontario**	**Total**	**897**
Canada	Québec	Black	11
Canada	Québec	Coulonge	17
Canada	Québec	Lièvre	263
Canada	**Quebec**	**Total**	**291**
Canada	**Total**	**Total**	**1361**
Total	**Total**	**Total**	**4550**

La quasi-totalité des systèmes hydriques présentés sont caractérisés par un régime hydrologique saisonnier dominé par l'accumulation et la fonte nivale. Les bassins versants régulés par les ouvrages de la Colombie Britannique sont caractérisés par des régions montagneuses de forts dénivelés, alors que ceux de l'Ontario, du Québec et du nord-est des États-Unis sont relativement plats. Ceci est également représentatif des régions étudiées dans le Mid-Ouest American et de la Californie, qui, quant à eux, se rapprochent d'un régime hydrologique mixte pluvial et nival à différents niveaux. En plus de présenter une grande variabilité géographique, ces rivières et les centrales qui y sont installées présentent une gamme de caractéristiques très variées tant en nombre de groupes installés qu'en

capacité de production totale. Certaines comportent un vaste réservoir de tête, d'autres sont au fil de l'eau, comme le montre le tableau 4.2.

Tableau 4.2 Caractéristiques globales des sites utilisés

	Nombre de groupes	Production moyenne (GWh)	Capacité de stockage (GWh)
Minimum	1	1	0
Maximum	16	1129	1095
Moyen	3	99	254

La variabilité temporelle de la production est très grande pour la plupart des rivières. La figure 4.2 en présente un exemple.

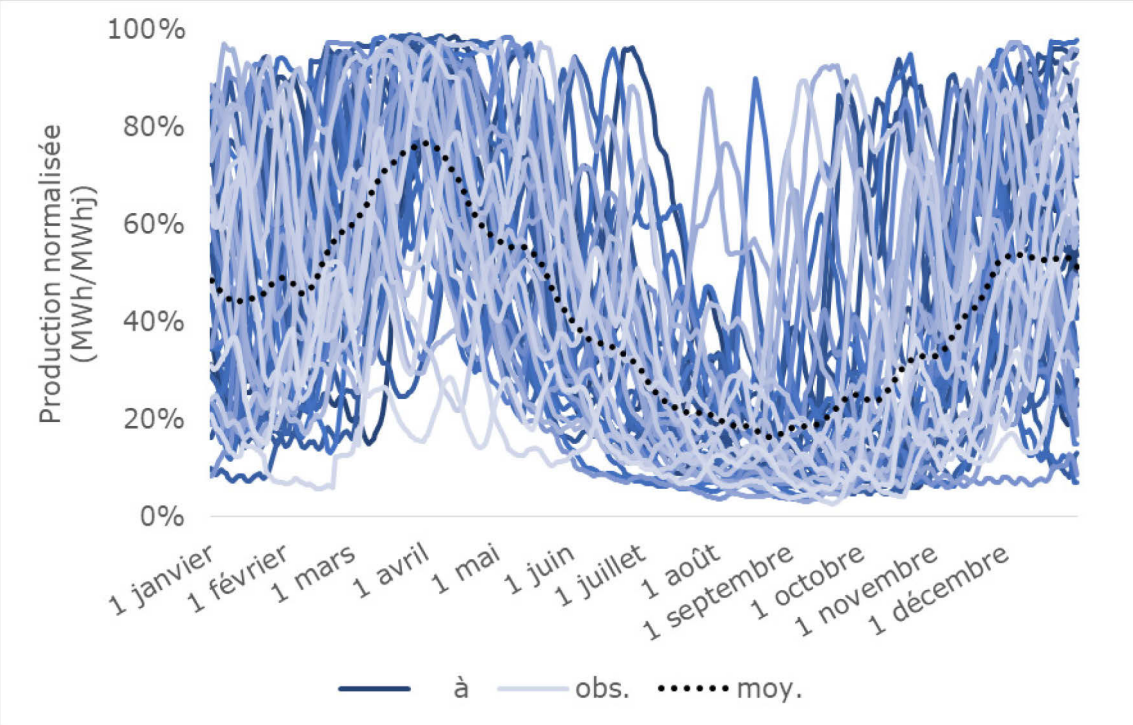


Figure 4.2 Exemple de données de production journalière

Chacun des traits bleus montre une année de production journalière pour une même rivière, normalisée en fonction de sa capacité installée. La moyenne est en pointillé noir. Sous forme cumulée, la figure 4.3 présente les volumes d'énergie disponibles pour la commercialisation et la planification des activités.

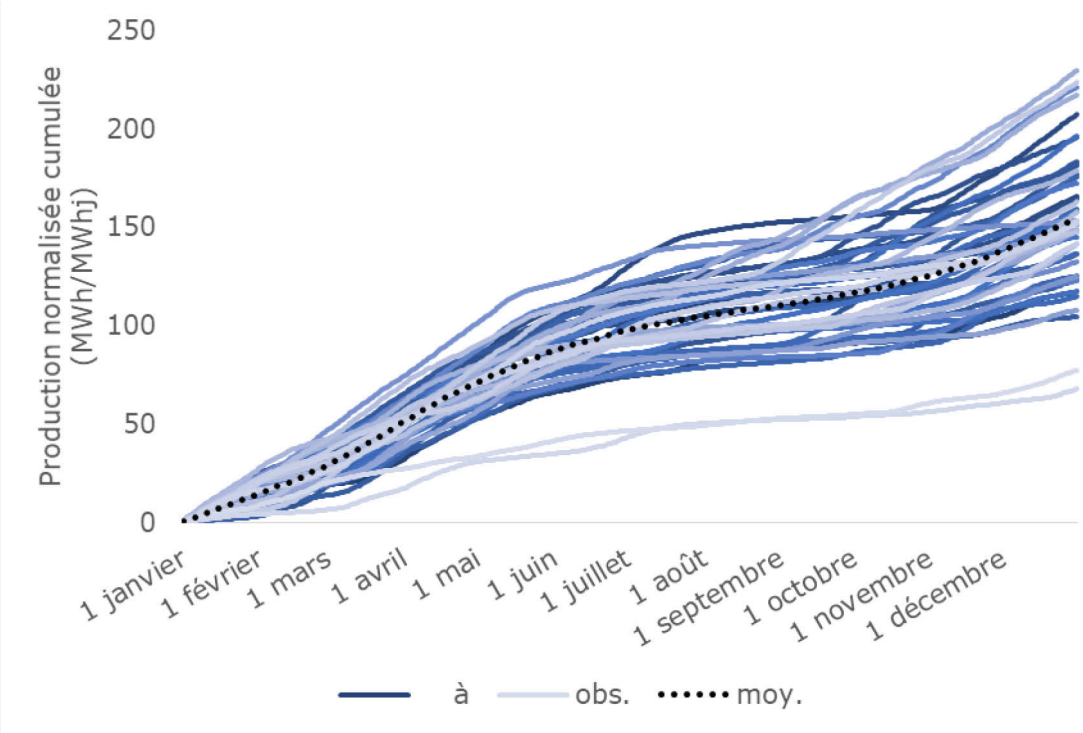


Figure 4.3 Exemple de données de production cumulée

La figure 4.3 montre les mêmes séries qu'à la figure 4.2, mais en cumulant les valeurs sur l'année. De cette façon, malgré une grande variabilité, les patrons de production cumulée suivent une tendance généralement semblable et la figure permet de voir que l'incertitude autour de la moyenne

sur les volumes produits augmente proportionnellement avec l'horizon observé.

Les séries de données de production journalière ont été fournies par Énergie Brookfield. Ces données sont considérées comme les meilleures disponibles. Pour chaque rivière, les séries s'étalent du 1er janvier 1969 au 31 décembre 2016. Ces séries ont été ajustées, pour représenter les capacités actuelles de chaque centrale, au fil des années par des firmes de consultants externes. La validation de ces ajustements est en dehors du cadre de cette recherche, ce point est traité plus en détail à la section travaux futurs. La centrale de la rivière Pit en Californie a été retirée du jeu de données. Cette centrale est la seule de la région du sud-ouest et sa production est influencée par plusieurs facteurs autres que l'hydrologie.

4.2 Application du modèle et discussion

Un jeu de paramètres initial, voir le tableau 4.3, a d'abord été imposé au modèle afin de servir de point de référence pour les tests d'optimisation présentés dans la prochaine section.

Tableau 4.3 Paramètres de référence

	k	***T***	***W***	***Début***	***Fin***
Paramètres	10	95	7	Sept.	Fév.

Ces paramètres sont utilisés pour tester le modèle. Tour à tour, chaque année devient la variable d'intérêt. Les autres années sont utilisées pour en faire la prévision. Ces résultats sont ensuite comparés aux valeurs réelles pour évaluer la performance du modèle. Selon ces paramètres, dix séries

analogues *k* seront sélectionnées selon leur ressemblance à la série d'intérêt dans les 95 jours *T* précédant le début de la prévision, en leur permettant un décalage maximal *W* d'une semaine. Pour ce banc d'essai, la date de début de prévision est prévue pour le premier septembre. Ainsi, la période *T* de comparaison entre les différentes années de production de rivières s'étire sur trois mois, débutant en juin. Les prévisions sont effectuées pour 150 jours jusqu'au début février de l'année suivante. Le RMSE obtenu, agrégé par région et normalisé selon la production observée, est présenté au tableau 4.4.

Tableau 4.4 RMSE relatif agrégé par région a) du modèle avec paramètres de références, b) en utilisant le LTA, c) l'amélioration en point de pourcentage

a) Horizon (jours)	10	30	90	150
Colombie Britannique	49%	30%	20%	17%
Midwest et sud-est	78%	35%	41%	35%
New York	72%	28%	34%	28%
Nouvelle Angleterre	44%	19%	21%	19%
Ontario	53%	22%	24%	22%
Québec	34%	17%	18%	17%

b) Horizon (jours)	10	30	90	150
Colombie Britannique	110%	59%	33%	26%
Midwest et sud-est	141%	86%	57%	46%
New York	146%	86%	54%	43%
Nouvelle Angleterre	116%	65%	39%	32%
Ontario	107%	59%	36%	30%
Québec	101%	55%	33%	27%

c) Horizon (jours)	10	30	90	150
Colombie Britannique	61%	29%	13%	9%
Midwest et sud-est	63%	50%	15%	11%
New York	74%	58%	19%	15%
Nouvelle Angleterre	72%	46%	18%	13%
Ontario	54%	36%	11%	8%
Québec	67%	38%	15%	10%

Ces résultats indiquent que malgré des paramètres non optimaux, la méthode proposée présente une nette amélioration par rapport à l'utilisation de la moyenne comme prévision. Les données de RMSE pour l'horizon de 10 jours montrent que le modèle est mieux adapté pour détecter les tendances se développant sur une longue période de temps plutôt que les variations à court terme. Quant à l'incertitude sur la prévision, les histogrammes de PIT permettent de qualifier la performance de la distribution bêta utilisée.

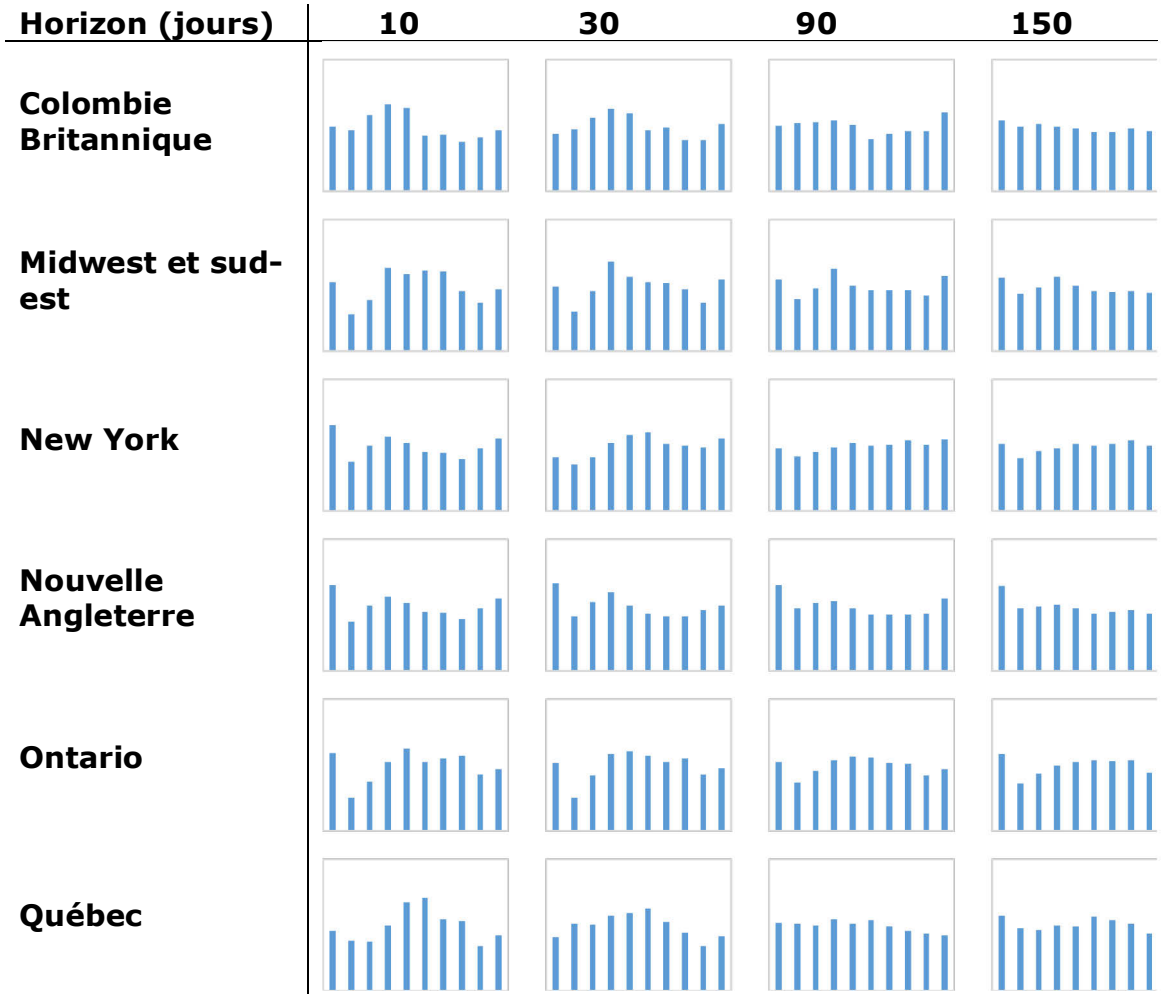


Figure 4.4 Histogrammes de PIT agrégés par région en utilisant les paramètres de références

Les histogrammes de la figure 4.4 des prévisions obtenues pour chacune des dix classes, sur une échelle de 0% à 20%, sont de plus en plus uniformes de la première à la quatrième colonne. Ce qui montre que la représentativité de la distribution bêta augmente avec le nombre de jours de prévisions. Ce qui peut s'expliquer par la plus grande taille d'échantillon.

4.2.1 Boucles d'optimisation des paramètres

Les résultats préliminaires obtenus avec des paramètres imposés sont encourageants et démontrent que le modèle est en mesure de fournir des prévisions de qualité et d'offrir une bonne représentativité de la distribution statistique des résultats. Afin de maximiser le potentiel de ce modèle, un jeu de paramètres *k, T* et *W* qui minimise le RMSE tout en conservant un histogramme de PIT doit être trouvé. Pour ce faire, le modèle est lancé de façon répétée à l'aide de boucles d'exécution imbriquées. L'utilisation des boucles imbriquées est une méthode dite naïve ce qui implique que d'autres façons de faire pourraient permettre de trouver un résultat optimal plus rapidement. Toutefois, étant donné l'ensemble de données relativement petit, la rapidité d'exécution de cette phase d'optimisation n'est pas un enjeu.

Tableau 4.5 Jeux de paramètres testés

Paramètres	***k***	***T***	***W***
Minimum	5	30	5
Maximum	40	360	30
Incrément	5	30	5

À l'aide de trois boucles imbriquées, plus de 600 itérations ont été nécessaires pour tester l'ensemble des combinaisons possibles de ces trois

paramètres. Les bornes ont été fixées de façon à ce que le nombre d'analogues soit assez élevé pour inclure presque chaque année d'historique disponible. Pour trouver ces séquences analogues, la longueur des séries peut s'étendre de un mois à un an et une déformation temporelle allant jusqu'à 30 jours est permise. Pour chaque rivière, toutes les itérations ont été effectuées et la combinaison de paramètres présentant le RMSE le plus bas sur l'horizon de 150 jours a été retenue. L'évolution des RMSE agrégés par région est montrée au tableau 4.6.

Tableau 4.6 RMSE agrégés par région

Horizon (150 jours)	Paramètres de référence	Paramètres optimaux
Colombie Britannique	17%	10%
Midwest et sud-est	35%	32%
New York	28%	21%
Nouvelle Angleterre	19%	17%
Ontario	22%	17%
Québec	17%	15%

L'optimisation des paramètres a permis d'améliorer le score RMSE de 2% à 7%. Ces résultats laissent à penser que l'algorithme est assez robuste et performe plutôt bien, malgré l'utilisation de paramètres sous-optimaux. Une analyse plus approfondie de ces résultats a permis de montrer qu'en effet, l'utilisation de paramètres sous-optimaux a une faible incidence sur la performance du modèle. En effet, le tableau 4.7 précise les fourchettes de paramètres optimaux et leur occurrence au fil des boucles d'optimisation.

Tableau 4.7 Fourchette de paramètres optimaux

Paramètres	*k*	*T*	*W*
Minimum	10	60	10
Maximum	20	90	15
Proportion	72%	65%	79%

4.2.2 Analyse saisonnière de la performance du modèle

Les résultats obtenus jusqu'à maintenant portent sur des prévisions débutant le 1er septembre. Étant donné le caractère saisonnier des données utilisées, l'impact de la date de début de prévision a été analysé en la déplaçant de l'automne au printemps, plus précisément au 1er mars. Ainsi, les effets de la fonte des neiges et des crues printanières seraient pris en compte par la période *T* de comparaison entre les séries analogues. Les résultats montrent une amélioration du RMSE en début d'horizon prévisionnel, mais que cet effet s'atténue, voir même s'inverse après 90 jours.

Tableau 4.8 Variation relative du RMSE selon la date de début de prévision

Horizon (jours)	10	30	90	150
Californie	6%	4%	2%	-12%
Colombie Britannique	6%	5%	1%	-6%
Midwest et sud-est	6%	5%	1%	-4%
New York	5%	4%	1%	-7%
Nouvelle Angleterre	8%	6%	1%	-3%
Ontario	6%	6%	1%	-1%
Québec	2%	4%	0%	-3%

Ces résultats indiquent que les années analogues au moment de la crue printanière sont moins représentatives sur une longue période de temps. La transition entre les périodes sèches et humides pourrait expliquer pourquoi les patrons de production observée sont moins corrélés sur un long horizon en débutant l'horizon au 1er mars. Ce constat est contrebalancé par une hausse de la précision sur les horizons à plus court terme. Par exemple, la production de deux années présentant des crues similaires peuvent diverger plus tard à l'automne. Sans s'y limiter, les trajectoires de production cumulée plus rectilignes causées par les volumes de crues importants pourraient être plus facilement prévisibles par l'algorithme. Les crues printanières engendrent une tendance lourde occasionnant moins de changement rapide de production.

4.2.3 Communication de l'incertitude

La représentation de prévision probabiliste peut être montrée par des quantiles et des moments de la distribution de probabilité comme la moyenne et la variance. Le graphique en éventail est une visualisation couramment utilisée qui représente un ensemble d'intervalles de prévision agrégés dans un graphique. Chacun des quantiles peut être interprété comme une évolution temporelle possible de la production.

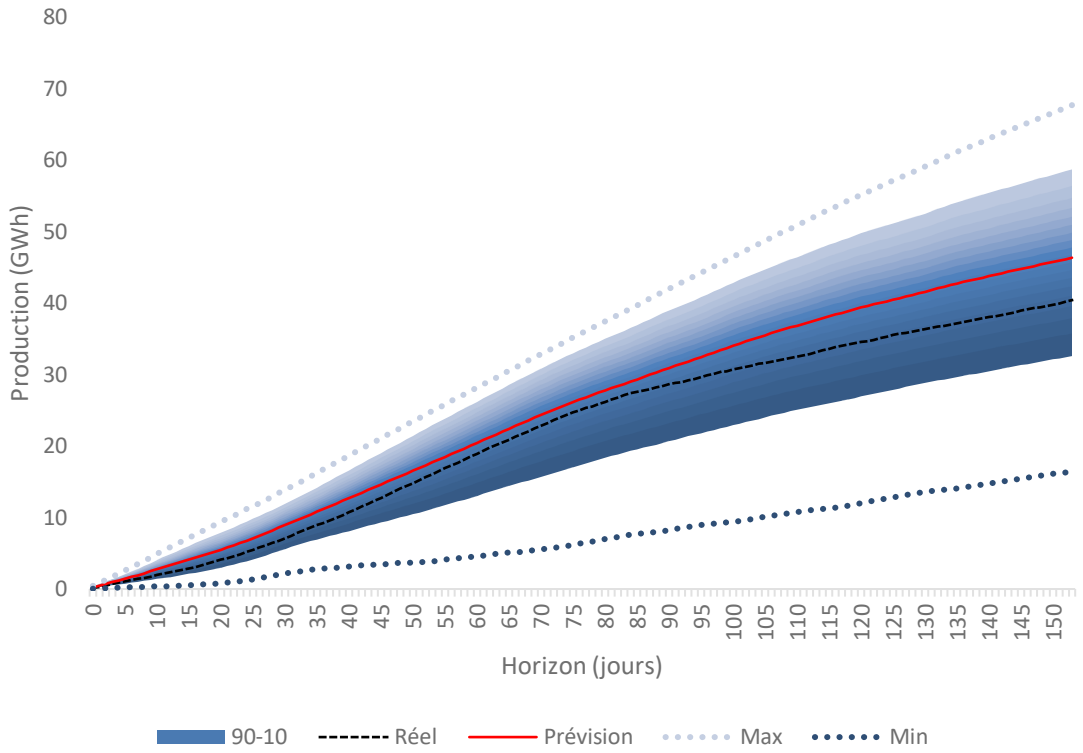


Figure 4.5 Exemple de prévision de la production au site Androscoggin

Le graphique en éventail de la figure 5 a été généré avec la méthode statistique décrite à la section 3.4. La transmission de l'information contenue dans ce graphique pourrait permettre au preneur de décision de se prémunir contre les risques d'approvisionnement en eau et de production d'électricité

5 CONCLUSION

Dans le cadre de cette recherche, le potentiel de produire des prévisions de production hydro-électrique par la seule utilisation des données historiques de production sous forme de séries temporelles a été exploré. L'utilisation d'un modèle statistique, basé sur l'historique, a été préconisée pour répondre à une certaine réalité du milieu. En effet, il est souvent souhaitable de limiter le volume de données à recueillir sur le terrain, inhérentes aux modèles à base physique. La réalité de plusieurs compagnies hydro-électriques fait en sorte que l'implantation de systèmes de prévisions complexes incluant des modèles de prévisions climatiques et hydrologiques servant d'intrants à des algorithmes d'optimisation de production électrique requiert une infrastructure informatique et des ressources techniques souvent trop grandes pour leur capacité. L'utilisation de données brutes permet de simplifier la modélisation.

Pour ce faire, les données de production de 181 centrales, réparties sur 68 systèmes hydriques au Canada et aux États-Unis ont été analysées. La conception du modèle a été basée sur différentes techniques de traitement de signal. De cette façon, l'hétérogénéité des caractéristiques physiques de ces installations (topographie, taille du bassin versant, occupation du sol, latitude, etc.) est intrinsèquement prise en compte par le modèle.

La méthode développée consiste à utiliser l'historique récent de production hydro-électrique, sous forme de série temporelle journalière, pour la comparer à l'ensemble des données disponibles à ce site. Les séquences similaires sont trouvées à l'aide d'une mesure de distance élastique. Cette

mesure permet de prendre en compte la variabilité naturelle des séries de production hydro-électrique et de maximiser l'alignement de la forme générale des séries. Les séquences similaires sont ensuite pondérées et leur production subséquente utilisée pour produire une prévision de la production future. L'incertitude liée à cette prévision est communiquéeau moyen d'une courbe de distribution cumulative bornée entre une production nulle et la production maximale au site sélectionné.

Cette méthodologie a été testée par simulation rétrospective, en l'appliquant en boucle sur l'historique et en comparant les prévisions obtenues avec les valeurs observées. Les résultats préliminaires de ces tests se sont montrés concluants, réduisant l'erreur quadratique moyenne obtenue par l'utilisation du LTA, un standard chez certains producteurs. L'analyse des histogrammes PIT a permis de constater visuellement que la qualité de la distribution augmentait avec l'horizon de prévision. Par la suite, à l'aide de boucles imbriquées, les principaux paramètres du modèle, soit la longueur de la série d'historique récent à utiliser T, la déformation autorisée w pour trouver les séquences similaires et le nombre K de ces séquences à utiliser, ont été optimisées.

Les objectifs principal et secondaire de ce projet ont été atteints. En effet, cette étude a montré la possibilité de produire un modèle de prévision basé sur les données historiques qui ne nécessite aucune instrumentation supplémentaire sur le terrain ni de systèmes complexes de prévisions. Ce modèle fournit une prévision de plus grande qualité que le LTA et permet de communiquer l'incertitude liée à ces prévisions de production.

Toutefois, le modèle présenté comporte plusieurs limites, en particulier l'hypothèse que l'historique est représentatif de la production future. Dans un contexte de changement climatique et d'évolution des conditions hydrologiques, les modes d'opérations des ouvrages et de production électriques sont portés à changer. Conséquemment, l'utilisation de ce modèle doit être encadrée et supervisée par un expert qui connaît la production historique de chaque site et qui est en mesure de porter un regard critique sur les résultats obtenus.

5.1 Perspectives de recherche et travaux futurs

La généralisation des conclusions tirées de cette étude devrait passer par son application à d'autres types de données. Sans s'y limiter, les domaines de la production éolienne, de la variation des prix de l'électricité et même de l'hydrologie directement pourraient être étudiés en variant le pas de temps utilisé et les différents paramètres. La modélisation de ces phénomènes est très complexe et tout comme pour la production hydro-électrique, la compréhension de chaque facteur influençant les fluctuations de vent, de prix ou de débit en rivière peut s'avérer une tâche ardue nécessitant plusieurs hypothèses ou une grande quantité de données prélevées sur le terrain. Ceci pourrait favoriser l'implantation d'un modèle basé sur l'historique, maximisant l'utilisation des données déjà disponibles.

De façon indirecte, l'application de ce modèle pourrait permettre de tirer certaines conclusions sur d'autres phénomènes dépassant le cadre de cette recherche. Ainsi, le support à la détection de tendance à long terme, comme les changements climatiques ou quel qu'autre phénomène, serait envisageable en analysant la distribution des séquences similaires. Si une cassure dans cette

distribution est observée, par exemple, si les années récentes sont plus fréquemment considérées comme analogues entre elles que les années de l'historique plus lointain, la stationnarité des données pourrait être remise en cause.

Également, les développements récents dans le domaine de l'intelligence artificielle sont prometteurs pour les modèles basés sur les données historiques. La combinaison de plusieurs de ces modèles serait envisageable afin d'améliorer la qualité des prévisions tout en limitant le besoin de nouvelles stations de mesures.

6 LISTE DES RÉFÉRENCES

1. Alsharif, M.H., Younes, M.K. et Kim, J., *Time series ARIMA model for prediction of daily and monthly average global solar radiation : The case study of Seoul, South Korea.* Symmetry, 2019. 11(2).

2. Anand, P. *Flood Hydrology, Proceeding of the International Symposium on Flood Frequency and Risk Analyses*. in *Flood Hydrology - Proceeding of the International Symposium on Flood Frequency and Risk Analyses*. 1987. Baton Rouge.

3. Azadzadeh, I., *Thèse - Hydrological Time Series Modelling and Applications*. 2016: Université de Calgary.

4. Beven, K., *A manifesto for the equifinality thesis.* Journal of hydrology, 2006. 320(12): p. 18-36.

5. Beven, K. et Binley, A.M., *The future of distributed models: model calibration and uncertainty prediction.* Hydrological process, 1992. 6: p. 279-298.

6. Box, G. et Jenkins, G., *Time Series Analysis: Forecasting and Control*. 1970, San-Francisco: Holden-Day.

7. Chambers, J.C., Mullick, S.K. et Smith, D.D., *How to Choose the Right Forecasting Technique.* Harvard business review, 1971. (Réimpression) 71403(Juillet-août): p. 45-69.

8. Chen, L., *Similarity search over time series and trajectory data*. 2005, Université de Waterloo.

9. Chen, M.S., Han, J. et Yu, P.S., *Data mining: an overview from a database perspective.* IEEE Transactions on knowledge and data engineering, 1996. 8(6): p. 866-883.

10. Cheng, C.T., Niu, W.J., Feng, Z.K., Shen, J.J. et Chau, K.W., *Daily reservoir runoff forecasting method using artificial neural network based on quantum-behaved particle swarm optimization.* Water, 2015. 7(8): p. 4232-4246.

11. Cobaner, M., Haktanir, T. et Kisi, O., *Prediction of Hydropower Energy Using ANN for the Feasibility of Hydropower Plant Installation to an Existing Irrigation Dam.* Water resources management, 2008. 22(6): p. 757-774.

12. Confias, J. et Le Guen, M., *Premiers pas en Régression Linéaire.* La Revue Modulad no 35, 2006: p. 260-277.

13. Côté, P., Haguma, D., Leconte, R. et Krau, S., *Stochastic optimisation of Hydro-Quebec hydropower installations: a statistical comparison between SDP and SSDP methods.* Canadian Journal of Civil Engineering, 2011. 38(12): p. 1427-1434.

14. Dau, A., Silva, D., Petitjean, F., Forestier, G., Bagnall, A. et Mueen, A., *Optimizing dynamic time warping's window width for time series data mining applications.* Data Mining and Knowledge Discovery, 2018. 32(5): p. 1-47.

15. Deng, J.S. et Oren, S.S., *Electricity derivatives and risk management.* Energy journal, 2006. 31(6): p. 940-953.

16. Diebold, F.X., Gunther, T.A. et Tay, A.S., *Evaluating density forecasts.* International economic review, 1998. 39: p. 863-882.

17. Énergie Brookfield. *Global presence*. [site web] 2019 [04/12/2019]; URL: https://renewableops.brookfield.com/en/presence/global-presence.

18. Esling, P. et Agon, C., *Time-Series Data Mining.* Computer Survey, 2012. 45(1): p. 12:1-12:34.

19. Frawley, W.J., Piatetsky-Shapiro, G. et Matheus, C.J., *Knowledge Discovery in Database: An Overview.* AI Magazine, 1992. 13(3): p. 57-70.

20. Fulcher, B.D. et Jones, N.S., *Highly comparative feature-based time-series classification.* IEEE Transactions in knowledge and data engineering, 2014. 28(12): p. 3026-3037.

21. Gençay, R. et Stengos, T., *Moving average rules, volume and the predictability of security returns with feedforward networks.* Journal of Forecasting, 1998: p. 401-414.

22. Goldin, D.Q. et Kanellakis, P.C., *On Similarity Queries for Time-Series Data: Constraint Specification and Implementation*. 1995: CP.

23. Grayson, R.B., Bloschl, G., Western, A.W. et McMahon, T.A., *Advances in the use of observed spatial patterns of catchment hydrological response.* Advances in water resources, 2002. 25: p. 1313-1334.

24. Guyon, I. et Elissef, A., *An introduction to variable and feature selection.* Journal of machine learning research, 2003. 3(3): p. 1157-1182.

25. Guyon, I., Weston, J., Barnhill, S. et Vapnik, V., *Gene selection for cancer classification using support vector machine.* Machine Learning, 2002. 46: p. 389-422.

26. Hassan, M.M. et Croke, B.F.W. *Filling gaps in daily rainfall data: a statistical approach*. in *20th International Congress on Modelling and Simulation*. 2013. Adelaide.

27. Helsel, D.R. et Hirsch, R.M., *Statistical Methods in Water Resources*, in *Techniques of water-resource Investigations of the United States Geological Survey, Book 4, Hydrologic Analysis and Interpretation*. 2002, USGS. p. 285-287.

28. Hjelmfelt, A. et Wang, M., *Artificial neural network as unit hydrograph applications.* 1993.

29. Hurst, H.E., *Long-Term Storage of Reservoirs: An Experimental Study.* Transactions of the American Society of Civil Engineers, 1951(116): p. 770-799.

30. Jimenez, C., Hipel, W. et McLeod, A., *Développements récents dans la modélisation de la persistance à long terme.* Revue des sciences de l'eau, 1990. 3: p. 55-81.

31. K. Ku-Mahamud, N.Z., N. Katuk and M. Shbier, *Flood Pattern Detection Using Sliding Window Technique.* Third Asia International Conference on Modelling & Simulation, 2009: p. 45-50.

32. Kang, K.W., Kim, J.H., Park, C.Y. et Ham, K.J. *Evaluation of hydrological forecasting system based on neural network model*. in *Proceeding of 25th Congress of International Association for Hydraulic Research*. 1993. Tokyo.

33. Kopp, S., Djovick, D. et Rea, A. *Introduction to GIS and Hydrology* in *ESRI International Preconference and Seminars*. 2005.

34. Koutsoyiannis, D., *Typical Distribution Functions in Geophysics, Hydrology and Water resources*, in *Probability and statistics for geophysical processes*. 2008, National Technical University of Athens: Athens. p. 1-33.

35. Koutsoyiannis, D. et Georgakakos, A.P. *Lessons from the long flow records of the Nile*. 2006. Rhodes.

36. Kumar, D.N., Raju, K.S. et Sathish , T., *River flow forecasting using recurrent neural network.* Water resources management, 2004. 18(2): p. 143-160.

37. Lall, U. et Sharma, A., *A nearest neighbor bootstrap for resampling hydrologic time series.* Water resources research, 1996. 32(3): p. 679-693.

38. Lane, D.M., *Introduction to linear regression*, in *Online Statistics Education: An Interactive Multimedia Course of Study*. 2007.

39. Langousis, A. et Koutsoyiannis, D., *A stochastic methodology for generation of seasonal time series reproducing overyear scaling.* Journal of hydrology, 2006(322): p. 3277-3284.

40. Levenshtein, V., *Binary codes capable of correcting deletions, insertions, and reversals.* Soviet physics doklady, 1966. 10(8): p. 707-710.

41. Lohani, A.K., Kumar, R. et Singh, R.D., *Hydrological time series modeling: A comparison between adaptive neuro-fuzzy, neural network and autoregressive techniques.* Journal of hydrology, 2012. 442: p. 23-35.

42. Madni, H.A., Anwar, Z. et Shah, M.A. *Data mining techniques and applications : A decade review*. in *23rd International Conference on Automation and Computing (ICAC)*. 2017. Huddersfield.

43. Minns, A.W. et Hall, M.J., *Artificial neural networks as rainfall-runoff models.* 1996. 41(3).

44. Mishra, S., Dwivedi, V.K., Saravanan, C. et Pathak, K.K., *Pattern Discovery in Hydrological Time Series Data Mining during the Monsoon Period of the High Flood Years in Brahmaputra River Basin.* International Journal of Computer Applications; Volume 67 - No. 6, 2013: p. 7-14.

45. Modrick, T. et Georgakakos, K., *Assessment of Folsom Lake response to historical and potential future climate scenarios.* Journal of hydrology, 2001. 249: p. 148-175.

46. Moffat, A.M., Papale, M., Reichstein, M. et Hollinge, M., *Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes.* Agricultural and forest meteorology, 2007: p. 209-232.

47. Mohan, A., *A New Spatio-Temporal Data Mining Method and*. 2014, University of Nebraska: Lincoln.

48. Mohanasundaram, S., Kumar, G.S. et Narasimhan, B., *A novel deseasonalized time series model with an improved seasonal estimate for groundwater level predictions.* H2O Open Journal, 2019. 2(1): p. 25-44.

49. Moradkhani, H., Hsu, K.L., Gupta, H. et Sorooshian, S., *Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter.* Water resources research, 2005. 41(5): p. 1-17.

50. Mori, U., Mendiburu, A. et Lozano, J.A., *Distance Measures for Time Series in R: The TSdist Package.* R Journal, 2016. 8(2): p. 451-459.

51. Moss, M.E. et Bryson, C.M., *Autocorrelation structure of monthly streamflows.* Water resources research, 1974: p. 737-747.

52. Neuman, S.P. *Accounting for conceptual model uncertainty via maximum likelihood Bayesian model averaging*. in *ModelCARE 2002*. 2002. Prague.

53. Noori, N. et Kalin, L., *Coupling SWAT and ANN models for enhanced daily streamflow.* Journal of hydrology, 2016. 533: p. 141-151.

54. Ouyang, R., Ren, L., Cheng, W. et Zhou, C., *Similarity search and pattern discovery in hydrological timeseries data mining.* Hydrological Processes 24, 2010: p. 1198-1210.

55. Perreault, L., *Vérification de prévisions hydrologiques probabilistes*. 2013, Hydro-Québec: Varennes.

56. Qin, G., Li, H., Wang, X., He, Q. et Li, S., *Annual runoff prediction using a nearest-neighbour method based on cosine angle distance for similarity estimation.* Remote Sensing and GIS for Hydrology and Water Resources, 2015. 368: p. 204-208.

57. Ratanamahatana, C.A. et Keogh, E., *Making time-series classification more accurate using learned constraints.* SDM International Conference, 2004: p. 11-22.

58. Sivakumar, B. et Berndtsson, R., *Advances in data-based approaches for hydrologic modeling and forecasting*. 2010, Singapour: World Scientific.

59. Sivapalan, M., Blosch, G., Merz, R. et Gutknecht, D., *Linking flood frequency to long-term water balance: Incorporating effects of seasonality.* Water resrouces research, 2005. 41(6).

60. Soltani, S., Modarres, R. et Eslamian, S.S., *The use of time series modeling for the determination of rainfall climates of Iran.* International Journal of Climatology, 2007. 27: p. 819-829.

61. Szolgayova, E., Laaha, G., Bloschl, G. et Bucher, C., *Factors inflencing long range dependence in dtreamflow of European rivers.* HYDROLOGICAL PROCESSES, 2014. 28: p. 1573-1586.

62. Tanty, R. et Desmukh, T.S., *Application of Artificial Neural Network in Hydrology - A Review.* International Journal of Engineering Research & Technology, 2015. 4(6): p. 184-188.

63. Thaeer Hammid, A., Herwan Sulaiman, M. et Abdalla, A., *Prediction of small hydropower plant power production in Himreen Lake dam (HLD) using artificial neural network.* Alexandria engineering journal, 2018. 57(1): p. 211-221.

64. Thirumalaiah, K. et Deo, M.C., *Hydrological Forecasting Using Neural Networks.* Journal of Hydrologic Engineering, 2000. 5(2).

65. Tibshirani, R., *Regression Shrinkage and selection via the Lasso.* Royal statistical society, Series B : Methodological, 1996. 58(1): p. 267-288.

66. Tsiporkova, E., *Center for plant system biology*. 2005.

67. Tyralis, H., Dimitriadis, P., Koutsoyiannis, D., O'Connell, P.E., Tzouka, K. et Iliopoulou, T., *On the long-range dependence properties of annual precipitation using a global network of instrumental measurements.* Advances in water resources, volume 3, 2018: p. 301-318.

68. Van del Dool, H., *Empirical Methods in Short-term Climate Prediction.* Oxford University Press, 2009. 175(1): p. 85-92.

69. Wagener, T. et Gupta, H.V., *Model identification for hydrological forecasting under uncertainty.* Stochastic environmental research and risk assessment, 2005. 19(6): p. 378-387.

70. Wang, K. et Gasser, T., *Alignment of Curves by Dynamic Time Warping.* The Annals of Statistics, 1997: p. 1251-1276.

71. Wang, L., Wang, X. et Ramamohanarao, K., *Characteristic based descriptors for motion sequence recognition.* Computer Science, 2008. 5012: p. 369-380.

72. Wasserman, L., *All of Statistics: A Concise Course in Statistical Inference*. 2005, Pittsburg: Springer.

73. Xing, Z., Pei, J. et Keogh, E., *A Brief Survey on Sequence Classification.* ACM SIGKDD Explorations Newsletter, 2010. 12(1): p. 40-48.

74. Yuan, P., Wang, W. et Ding, J., *Nonparametric perturbing nearest neighbor bootstrapping model for simulation of flood time series.* Journal of Sichuan University -Engineering Science Edition, 2000. 32(1): p. 82-86.